

IN THE UNITED STATES PATENT AND TRADEMARK OFFICE

In re Patent Application of:)
Yasuyuki FUJIKAWA)
Serial No.: To be assigned) Group Art Unit: Unassigned
Filed: January 4, 2001) Examiner: Unassigned



For: **STRUCTURAL DOCUMENTATION SYSTEM**

**SUBMISSION OF CERTIFIED COPY OF PRIOR FOREIGN
APPLICATION IN ACCORDANCE
WITH THE REQUIREMENTS OF 37 C.F.R. §1.55**

*Assistant Commissioner for Patents
Washington, D.C. 20231*

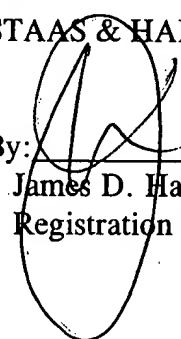
Sir:

In accordance with the provisions of 37 C.F.R. §1.55, the applicant submits herewith a certified copy of the following foreign application:

Japanese Patent Application No. 2000-027460
Filed: February 4, 2000.

It is respectfully requested that the applicant be given the benefit of the foreign filing date as evidenced by the certified papers attached hereto, in accordance with the requirements of 35 U.S.C. §119.

Respectfully submitted,
STAAS & HALSEY LLP

By: 
James D. Halsey, Jr.
Registration No. 22,729

700 11th Street, N.W., Ste. 500
Washington, D.C. 20001
(202) 434-1500
Date: January 4, 2001

日 本 国 特 許 庁
PATENT OFFICE
JAPANESE GOVERNMENT

10966 U.S. PTO
09/753514
01/04/01

別紙添付の書類に記載されている事項は下記の出願書類に記載されて
いる事項と同一であることを証明する。

This is to certify that the annexed is a true copy of the following application as filed
with this Office.

出 願 年 月 日
Date of Application:

2000年 2月 4日

出 願 番 号
Application Number:

特願2000-027460

出 願 人
Applicant(s):

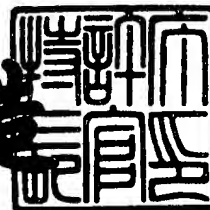
富士通株式会社

CERTIFIED COPY OF
PRIORITY DOCUMENT

2000年 8月18日

特 許 庁 長 官
Commissioner,
Patent Office

及 川 耕 造



出証番号 出証特2000-3064894

【書類名】 特許願

【整理番号】 9903156

【あて先】 特許庁長官殿

【国際特許分類】 G06F 13/368

【発明者】

【住所又は居所】 神奈川県川崎市中原区上小田中 4 丁目 1 番 1 号 富士通株式会社内

【氏名】 藤川 泰之

【特許出願人】

【識別番号】 000005223

【氏名又は名称】 富士通株式会社

【代理人】

【識別番号】 100098235

【弁理士】

【氏名又は名称】 金井 英幸

【手数料の表示】

【予納台帳番号】 062606

【納付金額】 21,000円

【提出物件の目録】

【物件名】 明細書 1

【物件名】 図面 1

【物件名】 要約書 1

【包括委任状番号】 9908696

【ブルーフの要否】 要

【書類名】 明細書

【発明の名称】 構造文書化システム

【特許請求の範囲】

【請求項 1】

テキスト形式で記述された処理対象の電子文書を、所定の文書構造を有する構造化文書に変換するための構造文書化システムであって、

前記文書構造を構成する基本単位である各要素間の相互関係を定義するとともに各要素毎にその抽出条件及び識別子を定義した定義情報を読み込む読込部と、

前記読込部によって読み込まれた定義情報によって定義された各要素毎の抽出条件を順次参照し、参照した要素の抽出条件に合致した領域を前記処理対象の電子文書から抽出する検索部と、

前記検索部によって各要素に関して抽出された領域を、前記定義情報によって定義された各要素間の相互関係に従って組み合わせるとともに、各領域に対して前記定義情報によって定義された識別子を付すことによって前記構造化文書を作成する構造化文書生成部と

を備えたことを特徴とする構造文書化システム。

【請求項 2】

前記構造化文書生成部は、前記識別子として、前記検索部によって抽出された各領域の前後にタグを付す

ことを特徴とする請求項 1 記載の構造文書化システム。

【請求項 3】

前記定義情報によって定義された各要素間の相互関係には、一つの上位階層の要素内に複数の下位階層の要素が含まれてなる階層構造が含まれ、

前記検索部は、前記上位階層の抽出条件に合致して抽出した領域に対して、前記各下位階層の要素の抽出条件を参照した抽出を行い、

前記構造化文書生成部は、下位階層を持たない末端の要素については前記検索部によって抽出された領域の前後にタグを付し、下位階層を持つ要素については当該要素の下位階層の全要素について前記検索部によって夫々抽出された領域を合わせてなる領域の前後にタグを付す

ことを特徴とする請求項 2 記載の構造文書化システム。

【請求項 4】

前記定義情報によって定義された各要素間の相互関係には、一つの上位階層の要素内に繰り返し構造を持つ下位階層の要素が含まれてなる階層構造が含まれ、

前記検索部は、前記上位階層の抽出条件に合致して抽出した領域に対して、繰り返し構造を持つ下位階層の要素の抽出条件を参照した抽出を、この抽出条件に合致する領域が抽出できなくなるまで繰り返し行い、

前記構造化文書生成部は、前記繰り返し構造を持つ下位階層の要素については前記検索部によって抽出された各領域の前後に夫々共通のタグを付すことを特徴とする請求項 3 記載の構造文書化システム。

【請求項 5】

前記定義情報によって定義された各要素間の相互関係には、一つの上位階層の要素内に抽出順序制約を持つ要素を含む複数の下位階層の要素が含まれてなる階層構造が含まれ、

前記検索部は、前記上位階層の抽出条件に合致して抽出した領域内において、前記抽出順序制約を持つ下位階層の要素の抽出条件を参照した抽出を、他の下位階層の要素の抽出条件に合致して既に抽出された領域の直後から行う

ことを特徴とする請求項 3 記載の構造文書化システム。

【請求項 6】

前記定義情報によって定義された何れかの要素の抽出条件は、抽出すべき領域全体の記述パターンである

ことを特徴とする請求項 1 記載の構造文書化システム。

【請求項 7】

前記定義情報によって定義された何れかの要素の抽出条件は、抽出すべき領域の開始部分の記述パターン及び終了部分の記述パターンである

ことを特徴とする請求項 1 記載の構造文書化システム。

【請求項 8】

前記記述パターンは、抽出すべき領域の文字列によって表現されていることを特徴とする請求項 6 又は 7 記載の構造文書化システム。

【請求項 9】

前記記述パターンは、抽出すべき領域の文字列に対応した正規表現によって表現されている

ことを特徴とする請求項 6 又は 7 記載の構造文書化システム。

【請求項 1 0】

前記定義情報によって定義された何れかの要素の抽出条件は、抽出すべき領域の構文要素である

ことを特徴とする請求項 1 記載の構造文書化システム。

【請求項 1 1】

コンピュータに対して、

テキスト形式で記述された処理対象の電子文書を読み込ませ、

構造化文書の文書構造を構成する基本単位である各要素間の相互関係を定義するとともに各要素毎にその抽出条件及び識別子を定義した定義情報を読み込ませ

読み込んだ定義情報によって定義された各要素毎の抽出条件を順次参照させ、参照した要素の抽出条件に合致した領域を前記処理対象の電子文書から抽出させ、

各要素の抽出条件に合致するものとして抽出された領域を、前記定義情報によって定義された各要素間の相互関係に従って組み合わせるとともに、各領域に対して前記定義情報によって定義された識別子を付加させることによって前記構造化文書を生成させる

プログラムを格納したコンピュータ可読媒体。

【発明の詳細な説明】

【0 0 0 1】

【発明の属する技術分野】

本発明は、テキストやソースプログラムリスト等の電子文書から、予め定義された所定の構造を有する領域を自動的に識別し、識別した領域に当該構造に対応したタグを付加することによって構造化文書に変換する構造文書化システムに、関する。

【 0 0 0 2 】

【従来の技術】

構造化文書とは、テキスト形式で記述された一般文書やソースプログラムリスト等の電子文書に、その電子文書内の各領域の意義（例えば、その領域の記述内容が当該電子文書のヘッダ部であるという事、その領域の記述内容が当該電子文書の作成日時を示す、その領域の記述内容が当該電子文書の作成者名であるという事、その領域の記述内容が閲覧ソフト上において拡大表示されるという事、等）を示すタグを付加したものである。その構造化文書には、そのタグの付加規約に従って、XML、SGML、HTMLといった形式がある。このうちXML、SGMLは、タグの種類をユーザが任意に設定できるものであり、XMLは、SGMLよりもその自由度を大きくしたものである。このような構造化文書において、各タグが付された領域相互の関係によって定義される電子文書の構造（例えば、ヘッダの次に本文が続き、ヘッダ内はタイトル、作成者、作成日時から構成される…といった構造）は、DTD（文書構造定義）と、呼ばれる。

【 0 0 0 3 】

図 3 3（a）は、XMLによる構造化文書の一例を示し、同図（b）は、このXML文書のDTD（文書構造定義）をツリー形式に表した図である。これら両図を照らし合わせることで理解されるように、DTDにおいて、構造化文章を構成する要素（一連の意義を持った領域）は、階層構造を有しており、それぞれに要素名（図 3 3（b）においては「報告書」、「ヘッダ」、「タイトル」…）が与えられている。即ち、最上位階層の要素「報告書」は、文書全体であって、要素「ヘッダ」及び複数の要素「内容」から構成されている。さらに、要素「ヘッダ」は、要素「タイトル」、要素「日付」、要素「担当者」及び要素「顧客名」から構成されている。そして、構造化文書のテキスト中における各要素の前後には、図 3 3（b）に示されるように各要素の要素名に対応したタグが付されている。例えば、要素「日付」の領域は、この要素名「日付」に対応したタグ<DATA>～</DATA>によって囲まれている。従って、XMLまたはSGML文書を扱うシステム（以下、「XML／SGMLシステム」という）は、このタグに囲まれている要素“1998.02.17”が日付を示しているものと、認識する。なお

、図33においては、説明を容易にするために、各要素名を日本語にて「報告書」、「ヘッダ」、「タイトル」…と表したが、実際には、XML及びSGMLの規約に従い、タグと同じアルファベットで表された名称となっていなければならない。

【0004】

このような構造化文書は、バイナリーファイルと異なり、基本的にテキストファイルであるが故にアプリケーションに依存しないという利点があるために、近年のインターネットの普及を背景に、インターネット等を通じて情報を交換したりデータベースにて管理するための文書形式として、広く用いられるようになって来ている。従って、このような構造化文書が定着する以前において作成された膨大な数量の電子文書を構造化文書に変換して、統一的に取り扱えるようにするという要請がある。従来、このような既存の電子文書の構造化文書への変換は、オペレータがエディタ画面上でその電子文書の内容を吟味して、DTDを参照しつつ、電子文書の意味内容に適したタグを手入力によって付加するしかなかった。

【0005】

一方、構造化文書の対象となる電子文書であるプログラムソースに関しては、従来から、BNF（バックス・ナウア・フォーム）に従った構文要素とコメントとの両方を解析して必要な情報を抽出するツールが、出始めている。しかしながら、従来のツールは、抽出できる内容や出力形式が固定であって、柔軟性がなかった。

【0006】

【発明が解決しようとする課題】

本発明の課題は、以上のような現状に鑑みて、テキスト形式で記述された処理対象電子文書に基づいて自動的に構造化文書を生成する構造化文書化システムの提供である。

【0007】

【課題を解決するための手段】

上記課題を解決するために、本発明による構造化文書システムは、テキスト形式

で記述された処理対象の電子文書を所定の文書構造を有する構造化文書に変換するために、前記文書構造を構成する基本単位である各要素間の相互関係を定義するとともに各要素毎にその抽出条件及び識別子を定義した定義情報を読み込む読込部と、前記読込部によって読み込まれた定義情報によって定義された各要素毎の抽出条件を順次参照し、参照した要素の抽出条件に合致した領域を前記処理対象の電子文書から抽出する検索部と、前記検索部によって各要素に関して抽出された領域を、前記定義情報によって定義された各要素間の相互関係に従って組み合わせるとともに、各領域に対して前記定義情報によって定義された識別子を付すことによって前記構造化文書を生成する構造化文書生成部とを、備えたことを特徴とする。

【 0 0 0 8 】

このように構成された本発明による構造化文書化システムによると、読込部によって読み込まれる定義情報には、変換の目標である構造化文書の文書構造を構成する各要素間の相互関係、及び、各要素に付されるべき識別子の他、各要素に対応する領域を処理対象電子文書から抽出する際の抽出条件が、定義されている。従って、検索部は、各要素毎にその抽出条件を順次参照することにより、各要素の抽出条件に合致した領域を、処理対象電子文書から抽出することができる。その結果、構造化文書生成部は、検索部によって抽出された領域を、定義情報によって定義された要素間の相互関係に従って組み合わせるとともに、各領域に対してその領域に対応した要素について定義情報によって定義された識別子を付すことによって、構造化文書として生成することができる。

【 0 0 0 9 】

本発明において処理対象とされる電子文書は、テキスト形式で記述されていさえすれば良いので、一般的な文章の他、J a v aソース等のソースプログラムリストも、これに含まれる。なお、ソースプログラムリスト内には、一般的な文章であるコメントが含まれていても良い。本発明において変換対象となる構造化文書、即ち、定義情報によって定義された識別子の種類は、XML形式によるものであっても良いし、SGML形式によるものであっても良い。これら各形式によるものである場合には、識別子は、各領域の前後に付されるタグである。

【 0 0 1 0 】

【発明の実施の形態】

以下、本発明による構造文書化システムの実施の形態を、図面を参照して説明する。

（実施形態の概略）

本発明による構造文書化システムは、図 2 に示すように互いにバス B によって接続された CPU 1，ハードディスク 2，RAM 3，ディスプレイ 4 及び入力装置 8 から構成される一般的なコンピュータシステムにおいて、ハードディスク 2 に格納されたプログラムを、CPU 1 によって RAM 3 上に読み出して、入力装置（キーボード及びマウス）8 を介して入力されるオペレータからの操作に従って順次 CPU 1 によって実行して、その処理結果をディスプレイ 4 上に表示することによって、実現される。即ち、ハードディスク 2 が、本発明におけるコンピュータ可読媒体に該当し、CPU 1 及び RAM 3 が、本発明における読込部、検索部、構造化文書生成部、コンピュータに、該当する。なお、図 2 においては、文章構造化システムを構成する全ハードウェアをローカルコンピュータのものと示したが、この構造文書化システムは、複数のコンピュータを LAN やインターネット等のネットワークを介して接続させてなる分散処理システムとして実現されても良い。

【 0 0 1 1 】

次に、以上のようにして実現される構造文書化システムの概略を説明する。図 1 は、本実施形態による構造文書化システム 5 の概念を、その拡張機能である DTD+パターン編集システム 6 及び DTD+パターン作成支援システム 7 の概念と共に示した概念図である。同図に示されるように、構造文書化システム 5 は、テキスト形式で記述された一般文書や BNF による文法に従って記述されたソースプログラムリスト等を、処理対象文書 T とする。また、この構造文書化システム 5 は、最終的に生成すべき構造化文書の文章構造（DTD）を構成する各要素間の相互関係を定義するとともに、この文章構造中の各要素に対応した領域を処理対象文書 T から自動的に抽出する際のキーとなる記述パターン（抽出条件）及びその領域に付すべきタグ（即ち、識別子としての要素名）を各要素毎に定義し

た「DTD+パターン情報（定義情報）」Rを、予め登録している。そして、「DTD+パターン情報」Rによって定義された各要素毎に、その要素の抽出条件に合致した領域を処理対象文書Tから抽出して、抽出した領域を「DTD+パターン情報」Rによって定義された各要素間の相互関係に従って組み合わせるとともに、「DTD+パターン情報」Rによって定義されているタグを各領域の前後に付す。このようにして、最終的には、各々タグが付された複数の領域からなる「構造化文書」Oを、生成して出力する。この「構造化文書」Oは、XML又はSGMLに従った構造を有しているので、一般のXML/SGMLシステムによって処理可能となる。

【 0 0 1 2 】

この「DTD+パターン情報」Rは、それ自体はテキスト形式で表されたファイルであるが、図7及び図12に示されるように、上述した一般のDTDと同様に、各要素の階層構造（即ち、一つの上位階層の要素内に複数の下位階層の要素が含まれてなる階層構造）を、ツリー形式で表すことができる。このようにツリー形式で表した場合に、最上位階層となる文書全体に対応した要素は「ルートノード」と称される。また、対象要素の直下の階層に存在する要素は当該対象要素にとっての「子ノード」と称され、逆に、対象要素の直上の階層に存在する要素は当該対象要素にとっての「親ノード」と称される。さらに、「子ノード」の「子ノード」は「孫ノード」と称される。また、同じ「親ノード」の「子ノード」の中では、ツリー図における上方に存在しているものが、下方に存在しているものにとっての「兄ノード」と称され、下方に存在しているものが、上方に存在しているものにとっての「弟ノード」と称される。特に、同じ「親ノード」の「子ノード」のうち最上方に存在しているものは、その「親ノード」にとっての「長子ノード」と称される。なお、同じ要素名の要素が（即ち、同じ構造の要素）が繰り返す場合には、その要素名には“*”が付されることによって、『繰り返しあり（繰返構造）』の旨が指定される。

【 0 0 1 3 】

但し、この「DTD+パターン情報」Rは、各要素毎にその抽出条件として記述パターンが定義されている点において、一般のDTDと異なる。この記述パタ

ーンの指定の仕方としては、抽出すべき領域の開始パターン及び終了パターンを文字列そのもの又は正規表現によって指定する仕方と、抽出すべき領域の全域を正規表現（そのルールの一部を図11に示す）によって指定する仕方とが、利用可能である。前者の場合には、さらに、その開始パターン又は終了パターンを抽出すべき領域内に含むか、開始パターンの直後以降を抽出すべき領域とするか、終了パターンの直前までを抽出すべき領域とするかを、指定することが可能である。これらの様々な指定の仕方は、同一の「DTD+パターン情報」R内に混在していても良くなっている。なお、処理対象文書TがBNF（バックス・ナウア・フォーム）による文法（そのルールの一部を図14に示す）に従って記述されたソースプログラムリストである場合には、後述するように、BNFに従った「構文要素（シンタックス）」自体によって指定する仕方が、利用される。この場合には、「構文要素」の前後に存在するコメントをも合わせて抽出する旨を指定することも可能であり、さらに、このコメント部分についての子ノードに関しては、上述した文字列そのもの又は正規表現によって記述パターンを指定する仕方が、採用される。何れの場合においても、処理対象文書Tの全体を示す要素の抽出条件は、特殊な場合として、「文書（テキスト）全体」と指定される。以上のような記述パターンを様々な指定する「DTD+パターン情報」R内の情報を、以下、記述パターン情報という。

【0014】

図7は、図8に示すような一般のテキストが処理対象文書Tである場合に適用される「DTD+パターン情報」Rの例を、ツリー構造で示した図である。この図7に示される例において、要素「ヘッダ」を抽出するための記述パターンは、“タイトル：”という文字列自体からなる記述パターンの直後から、『行頭から0個以上の空白の後“3.”という文字列があってその後に任意の文字が続く』記述パターンの直前までの領域が、抽出対象領域であることを示す。また、要素「ヘッダ」の子ノードである要素「日付」を抽出するための記述パターンは、要素「ヘッダ」の記述パターンによって抽出される領域中において、“対応日：”という文字列自体からなる記述パターンの直後から、その後最初の改行の直前までの領域が、抽出対象領域であることを示す。また、「繰り返しあり」の旨が指

定された要素「内容」を抽出するための記述パターンは、『行頭から0個以上の空白の後“4”乃至“9”の何れか及び“.”からなる文字列があってその後任意の文字が繰り返した後改行する』記述パターンの直後から、『行頭の直後に改行する』記述パターンの直前までの領域が、抽出対象領域であることを示す。

【0015】

図7に示される「DTD+パターン情報」Rを参照して図8に示される処理対象文書Tに対する処理を行った場合に、抽出された領域が「DTD+パターン情報」Rによって定義された相互関係に従って階層化されてなるツリー構造を、図9に示す。このツリー構造において、上記要素「ヘッダ」として抽出された領域は“商談対応報告書～1997.02.17”であり、上記要素「日付」として抽出された領域は“1997.02.17”であり、上記要素「内容」として抽出された領域は、“以下のSDAS～YPS”及び“当社説明員による～回答も行う”の二つである。更に、図9に示されるツリー構造に基づいて、各要素に対応して抽出された領域の前後にその要素名をタグとして付すことによって作成された構造化文書Oを、図10に示す。

【0016】

図12は、図13に示されるようなソースプログラムリスト（より具体的には、Javaソース）が処理対象文書Tである場合に適用される「DTD+パターン情報」Rの例を、ツリー構造で示した図である。なお、ソースプログラムリストが処理対象文書Tである場合には、構造化文書化システム5は、この処理対象文書Tに含まれる各構文要素の範囲及び内容を、図14にその一部が示されるBNF（バックス・ナウア・フォーム）を定義した構文分解定義ファイルBに従って、図15に示すように解析する。そして、解析された構文要素がなす階層構造を、図16に示されるようなツリー構造（構文木・コメント木）として、RAM3上に構築する。図14乃至図16から明らかなように、BNFによると、例えば、ClassDefinitionには、Name（図13、図15の例では“顧客”）及び、MethodDefinition又はFieldDefinitionが含まれ、MethodDefinitionにも、Name（図13、図15の例では“信用ランク”）が含まれる。

【0017】

図 1 2 に示される「DTD+パターン情報」R では、要素「クラス仕様」を抽出するための記述パターン“Comment+ClassDefinition”は、BNFに定義された構文要素「ClassDefinition」に合致した領域、及び、この領域の直前に連続したコメントの領域が、抽出対象領域であることを示す。また、要素「クラス仕様」の子ノードである要素「作成者」を抽出するための記述パターンは、要素「クラス仕様」の記述パターンによって抽出されるコメント領域中において、“作成者”という文字列自体からなる記述パターンの直後から、その後最初の改行の直前までの領域が、抽出対象領域であることを示す。また、要素「クラス仕様」の子ノードである要素「クラス名」を抽出するための記述パターン“ClassDefinition.Name”は、要素「クラス仕様」の記述パターンによって抽出される構文要素領域中において、BNFに定義された構文要素「Name」に合致した領域が、抽出対象領域であることを示す。また、また、要素「クラス仕様」の子ノードである要素「メソッド仕様」を抽出するための記述パターン“Comment+MethodDefinition”は、要素「クラス仕様」の記述パターンによって抽出される構文要素領域中において、BNFに定義された構文要素「MethodDefinition」に合致した領域、及び、この領域の直前に連続したコメントの領域が、抽出対象領域であることを示す。また、要素「メソッド仕様」の子ノードである要素「メソッド名」を抽出するための記述パターン“MethodDefinition.Name”は、要素「メソッド仕様」の記述パターンによって抽出される構文要素領域中において、BNFに定義された構文要素「Name」に合致した領域が、抽出対象領域であることを示す。また、要素「クラス仕様」の子ノードである要素「説明」を抽出するための記述パターンは、要素「メソッド仕様」の記述パターンによって抽出されるコメント領域中において、“説明：”という文字列自体からなる記述パターンの直後から、その後の改行以外の任意の文字列からなる領域が、抽出対象領域であることを示す。また、要素「メソッド仕様」の子ノードである繰返構造を指定された要素「パラメタ」を抽出するための記述パターン“MethodDefinition.PrameterName”は、要素「メソッド仕様」の記述パターンによって抽出される構文要素領域中において、BNFに定義された構文要素「Prameter」に合致した全ての領域が、抽出対象領域であることを示す。

【0018】

図12に示される「DTD+パターン情報」Rを参照して図13に示される処理対象文書Tに対する処理を行った場合に、抽出された領域が「DTD+パターン情報」Rによって定義された相互関係に沿って階層化されてなるツリー構造を、図17に示す。このツリー構造において、上記要素「クラス仕様」として抽出される領域は

```

"/** COPYRIGHT Fujitsu LTD
* 作成者  藤川  泰之（富士通）
* 更新者  原田  義之（富士通）
* 更新者  和田  憲明（富士通）
*/
public class 顧客 {
/*
*説明：資本金から信用度を割り出す。
*/

    public String 信用ランク(
        int 現在借入金
        long 公定歩合)
    {
        :
        :
    }

//説明：資本金。

```

【0019】

```

    public static int 資本金:
} "

```

である。

【0020】

また、上記要素「作成者」として抽出される領域は“* 作成者 藤川 泰之（

富士通) ” であり、上記要素「クラス名」として抽出される領域は“顧客” であり、上記要素「メソッド仕様」として抽出される領域は

“/*

*説明：資本金から信用度を割り出す。

*/

public String 信用ランク(

int 現在借入金

long 公定歩合)

{

:

:

} ”

である。

【0021】

また、上記要素「メソッド名」として抽出される領域は“信用ランク” であり、上記要素「説明」として抽出される領域は、“資本金から信用度を割り出す。” であり、上記要素「パラメタ」として抽出される領域は、“int 現在借入金” 及び“long 公定歩合” の二つである。

【0022】

図1に戻り、以上のようにして構造文書化システム5によって参照される「DTD+パターン情報」Rは、DTD+パターン編集システム6によって編集される。このDTD+パターン編集システム6は、図18に示すようなGUI（編集画面）を有するテキストエディタである。このDTD+パターン編集システム6の編集画面の左半分はDTDツリー構造リストボックス61となっており、その右半分は項目入力部62となっている。また、この画面の下縁近傍には、削除ボタン63、キャンセルボタン64、更新終了ボタン65、内容反映ボタン66、子として追加ボタン67、及び、弟として追加ボタン68が、並べて表示されている。

【0023】

D T D ツリー構造リストボックス 6 1 は、編集中的「D T D + パターン情報」R によって定義される各要素の要素名を、各要素間の階層構造を示すツリー図として表示するリストボックスである。この D T D ツリー構造リストボックス 6 1 に表示されている何れかの要素名が入力装置（マウス）8 を介してオペレータによってクリックされると、そのクリックされた要素名が示す要素が処理対象として選択され、その表示色が変更されるとともに（図 1 8 の例においては要素名「タイトル」の表示色が変更されている）、その要素名が示す要素についての現在の設定内容が、項目入力部 6 2 の各テキストボックスやチェックボックスやオプションボタンに表示される。

【0024】

この項目入力部 6 2 は、要素名テキストボックス 6 2 1、繰り返しありチェックボックス 6 2 1 0、パターンの意味オプションボタン 6 2 2、前後空白除去チェックボックス 6 2 2 0、パターン／開始パターン指定部 6 2 4、終了パターン指定部 6 2 5、親に対する範囲制限オプションボタン 6 2 6、出力タグ名テキストボックス 6 2 7 を含む。

【0025】

要素名テキストボックス 6 2 1 は、現在選択されている要素の要素名を表示する（記述される）テキストボックスである。また、繰り返しありチェックボックス 6 2 1 8 は、現在選択されている要素に繰り返しあり（繰返構造）とするか否かを表示する（指定される）するチェックボックスである。また、パターンの意味オプションボタン 6 2 2 は、現在選択されている要素における記述パターンの指定の仕方が、要素の開始パターン及び終了パターンを指定する仕方であるか要素全体の記述パターンそのものを指定する仕方であるかを表示する（指定される）オプションボタンである。また、前後空白除去チェックボックス 6 2 2 0 は、現在選択されている要素を抽出する際に抽出対象領域の前後に空白が含まれていたならばこれを除去するか否かを表示する（指定される）チェックボックスである。行頭文字除去テキストボックス 6 2 3 は、現在選択されている要素を抽出する際に抽出対象領域の行頭に含まれていたならば削除される文字列を表示する（記述される）テキストボックスである。

【0026】

パターン／開始パターン指定部624は、現在選択されている要素全体の記述パターン（パターンの意味オプションボタン622においてパターンそのものが指定されている場合）又は開始パターン（パターンの意味オプションボタン622において開始と終了が指定されている場合）の内容を表示する（指定される）領域である。このパターン／開始パターン指定部624は、パターンタイプオプションボタン6241、コメント処理チェックボックス領域6242、内容にパターンを含むチェックボックス6243、構文要素名参照ボタン6244、パターン記述テキストボックス6245を含む。

【0027】

パターンタイプオプションボタン6241は、対象記述パターンが文字列そのものであるか正規表現であるか構文要素名があるかを表示する（指定される）オプションボタンである。また、コメント処理チェックボックス領域6242は、パターンタイプオプションボタン6241にて構文要素名が選択されている場合において構文要素の前方に連続するコメントを抽出するか否かを表示する（指定される）チェックボックス、及び、同様に構文要素の後方に連続するコメントを抽出するか否かを表示する（指定される）チェックボックスを含む領域である。内容にパターンを含むチェックボックス6243は、パターンの意味オプションボタン622にて開始と終了が選択されている場合において、記述パターンに対応した文字列を抽出対象領域に含むか否かを表示する（指定される）チェックボックスである。また、構文要素名参照ボタン6244は、パターンタイプオプションボタン6241にて構文要素名が選択されている場合において、構文分解定義ファイルBに定義されている各構文要素名及びその定義内容のリストを表示するためにクリックされるボタンである。また、パターン記述テキストボックス6245は、現在選択されている要素における全体の記述パターン（パターンの意味オプションボタン622においてパターンそのものが指定されている場合）又は開始パターン（パターンの意味オプションボタン622において開始と終了が指定されている場合）自体を表示する（記述される）テキストボックスである。

【0028】

終了パターン指定部 6 2 5 は、現在選択されている要素における終了パターン（パターンの意味オプションボタン 6 2 2 において開始と終了が指定されている場合）の内容を表示する（指定される）領域である。この終了パターン指定部 6 2 5 は、パターンタイプオプションボタン 6 2 5 1、内容にパターンを含むチェックボックス 6 2 5 5、構文要素名参照ボタン 6 2 5 4、パターン記述テキストボックス 6 2 5 5 を含む。これらの機能は、パターン／開始パターン指定部 6 2 4 のものと全く同じであるので、その説明を省略する。

【 0 0 2 9 】

親に対する範囲制限オプションボタン 6 2 6 は、現在選択されている要素の親ノードに指定された記述パターンが構文要素である場合、当該要素の検索範囲が親ノードに対応する領域の全体であるか（無し）、親ノードに対応する領域における構文要素の部分であるか（構文要素）、親ノードに対応する領域における構文要素の前方に連続したコメント部であるか（前方コメント）親ノードに対応する領域における構文要素の後方に連続したコメント部であるか（後方コメント）を表示する（指定される）オプションボタンである。

【 0 0 3 0 】

出力タグ名テキストボックス 6 2 7 は、現在選択されている要素に対応する領域が抽出された後にその領域の前後に付加されるタグ（通常は要素名テキストボックス 6 2 1 に表示された要素名と同じ）を表示する（記述される）テキストボックスである。

【 0 0 3 1 】

何れかの要素が選択された状態において、オペレータが削除ボタン 6 3 をクリックすると、その要素の設定内容（DTD 構造及び記述パターン情報）が消去される（この場合には、項目入力部 6 2 内の各テキストボックス、チェックボックス及びオプションボタンは、全て空欄となる。）。

【 0 0 3 2 】

また、何れかの要素が選択された状態において、オペレータがキャンセルボタン 6 4 をクリックすると、その要素の選択が解除される（この場合には、項目入力部 6 2 内の各テキストボックス及びオプションボタンが全て空欄となるとも

に、DTDツリー構造リストボックス61内における当該要素の要素名の表示色が元に戻る。）。

【0033】

また、何れかの要素が選択された状態において、オペレータが項目入力部62内の何れかのテキストボックス内の記述又は何れかのチェックボックス又はオプションボタンにおけるチェック内容を変更した後に、内容反映ボタン66をクリックすると、その要素の設定内容が、項目入力部62に現在表示されている内容に変更される。

【0034】

また、何れかの要素が選択された状態において、オペレータが項目入力部63内における少なくとも要素名テキストボックス621の記述を変更した後に、子として追加ボタン67をクリックすると、その要素の子ノードとして、項目入力部62に現在表示されている設定内容を有する要素が、追加される。

【0035】

また、何れかの要素が選択された状態において、オペレータが項目入力部63内における少なくとも要素名テキストボックス621の記述を変更した後に、弟として追加ボタン68をクリックすると、その要素の弟ノードとして、項目入力部62に現在表示されている設定内容を有する要素が、追加される。

【0036】

さらに、DTDツリー構造リストボックス61内に表示されている要素名が入力装置8を介してオペレータによってドラッグされて、他の何れかの要素名上にドロップされると、ドラッグされた要素名が示す要素の設定内容が、ドロップ先の要素名が示す要素の子ノードとなるように、変更される。

【0037】

最後に、オペレータが内容反映ボタン66をクリックすると、現在における各要素の設定内容に従って、DTDパターン情報Rが作成又は更新される。

【0038】

以上のような編集画面及びこの編集画面に関連付けられた機能を有するDTD+パターン編集システム6を用いて、オペレータは、自在にDTD+パターン情

報 R を編集することができる。オペレータは、この DTD+パターン編集システム 6 を用いて、ゼロから DTD+パターン情報 R を作成することもできるが、図 1 に示した DTD+パターン作成支援システム 7 によって作成された DTD+パターン情報 R を、DTD+パターン編集システム 6 を用いて編集することによって完成させても良い。

【 0 0 3 9 】

この DTD パターン作成支援システム 7 は、図 2 4 に示すような GUI（選択画面）を有するテキストエディタである。この DTD パターン作成支援システム 7 は、図 2 5 及び図 2 6 に示すような複数の典型パターン定義情報 S を有している。各典型パターン定義情報 S は、定型的な文書に頻繁に表れる典型的な文字列パターン（以下、「典型パターン」という）を要素として抽出するための記述パターン情報の雛形を、定義している情報である。即ち、図 2 5、図 2 6 に示すように、各典型パターン定義情報 S は、その典型パターンの概略構造を特定する構造特定部 S 1、構造特定部 S 1 中において典型パターンの概略構造を構成する個々のエレメント（隅付き括弧で囲まれた部分）として用いられ得る文字の種類を正規表現で特定する文字種特定部 S 2、及び、DTD パターン情報 R における各要素毎の記述パターン情報の雛形である雛形部 S 3 から、構成されている。

【 0 0 4 0 】

図 2 5 の例は、例えば、

“会社名：富士通株式会社”

の様に、『行頭から 0 個以上の空白の後に、項目名（タイトル）、区切り文字（区切り）、具体的内容（内容）と続き、行末に到る』という典型パターンを一つの要素として抽出するための記述パターン情報の雛形を定義しているので、構造特定部 S 1 には、“《行頭》＊[タイトルパターン（項目名に対応）]＊[区切りパターン（区切り文字に対応）]＊[内容パターン（具体的内容に対応）]＊《行末》”と、特定されている。また、文字種特定部 S 2 では、[タイトルパターン] 及び [内容パターン] については“《改行以外》+”と特定され、[区切りパターン] については“；：／（）”と特定されている。また、雛形部 S 3 では、パターン指定方式が“開始と終了”と特定され、開始パターンが正規表現に

よる“《行頭》＊[タイトル文字列1] | [タイトル文字列2] ＊[区切り文字列1] | [区切り文字列2] ＊”と特定され、終了パターンが正規表現による“＊《行末》”と特定されている。[タイトル文字列1]及び[タイトル文字列2]は、項目名として採り得る記述候補が代入される部分である。同様に、[区切り文字列1]及び[区切り文字列2]は、区切り文字列として採り得る記述候補が代入される部分である。

【0041】

図26の例は、一つの親ノード及び複数の子ノードとして抽出される典型パターンに用いられる典型パターン定義情報Sであるので、雛形部S3としては、親ノードを抽出するための記述パターン情報の雛形（以下、「親ノード雛形S3a」という）と、構造特定部S1に記載された各[タイトルパターン]に夫々対応した子ノードを夫々抽出するための複数の記述パターン情報の雛形（以下、「子ノード雛形S3b」という）とが、備えられている。従って、親ノード雛形S3aには、各子ノードの要素名が夫々代入される[タイトルパターン1]～[タイトルパターン5]が、含まれている。また、各子ノード雛形S3bでは、兄ノードとの関係が、“順序性=有り”と特定されている。

【0042】

図24に示す選択画面は、ルート文書要素名テキストボックス71，サンプルプレビューボックス72，ツリーリストボックス73，及び、典型パターン選択領域74を、含んでいる。この典型パターン選択領域74は、各典型パターン定義情報Sに対応付けられた複数の典型パターン選択ボタン741を含んでいる。そして、各典型パターン選択ボタン741の表面には、それに対応付けられた典型パターン定義情報Sにおける構造特定部S1の特定内容を端的に示す文字列が、表示されている。例えば、図25に示される典型パターン定義情報Sは、図24における一番上の典型パターン選択ボタン741に対応付けられているので、この典型パターン選択ボタン741には、文字列“タイトル：NNNNNNNNN”と、表示されている。

【0043】

DTD+パターン作成支援システム7は、処理対象文書Tのサンプルを読み込

むと、そのテキスト内容を、サンプル表示ダイアログ 7 2 内に表示する。ルート文字要素名テキストボックス 7 1 は、サンプル表示ダイアログ 7 2 内に表示されたサンプルに基づいて作成される DTD+パターン情報 R におけるルートノードの要素名が記述されるテキストボックスである。このルート文字要素名テキストボックス 7 1 内にルートノードの要素名がオペレータによって書き込まれることにより、DTD+パターン作成支援システム 7 は、このルートノードの要素名のみを原始内容とする DTD+パターン情報 R を生成する。

【 0 0 4 4 】

DTD+パターン作成支援システム 7 は、サンプル表示ダイアログ 7 2 内に表示されたテキストにおける何れかの行がドラッグされて選択された後に、何れかの典型パターン選択ボタン 7 4 1 がクリックされると、その典型パターン選択ボタン 7 4 1 に対応付けられた典型パターン定義情報 S を読み出し、その構造特定部 S 1 内で特定された典型パターンの概略構造を選択された行に当てはめて、その概略構造を構成する各エレメントに相当する文字列を抽出する。そして、抽出した各エレメントに関する文字列の文字種を、文字種特定部 S 2 にて特定された文字種に変換する。そして、変換後における各エレメントに関する文字列を、雛形部 S 3 の記述パターン情報の雛形における各 [] に、代入する。このようにして、DTD+パターン作成支援システム 7 は、ルート文書要素名テキストボックス 7 1 に記述された要素名を有するルートノードの子ノード（若しくは、子ノード及び孫ノード）を抽出するための記述パターン情報を作成し、この指定内容を DTD+パターン情報 R に追加する。

【 0 0 4 5 】

ツリーリストボックス 7 3 は、作成中の DTD+パターン情報 R に含まれる各要素の要素名がその階層構造を示すツリー図として表示されるリストボックスである。従って、オペレータが、サンプル表示ダイアログ 7 2 内に表示されているテキスト中における何れか行をドラッグして何れかの典型パターン選択ボタン 7 4 1 をクリックする毎に、ツリーリストボックス 7 3 中に表示されているルートノードの下位階層に、子ノード（若しくは、子ノード及び孫ノード）の要素名が追加される。

(構造文書化システムの詳細構成及び処理内容)

次に、上述した構造文書化システム5の詳細な構成を、その処理内容に沿って説明する。図3は、この構造文書化システム5の詳細な構成(構造文書化システム5を構築するプログラムの詳細なモジュール構成)を示すブロック図である。また、図4乃至図6は、この構造文書化システム5の処理内容(構造文書化システム5を構築するプログラムによるCPU1の処理内容)を示すフローチャートである。

【0046】

図3に示されるように、構造文書化システム5は、DTD+パターンツリー作成部51、全体コントロール部52、パターン検索部53、及び、構文木分解部54から、構成されている。更に、パターン検索部53は、文字列検索部531、正規表現検索部532及び構文要素検索部533を、含んでいる。

【0047】

構文木分解部54は、処理対象文書Tがソースプログラムリスト(但し、BNFによる文法に従って記述されていることが条件)である場合に起動し、この処理対象文書Tの内容を構文分解定義ファイルBに従って解析して、解析した処理対象文書Tの構文構造に応じて、図16に示したような構文木・コメント木57をRAM3上に構築する。

【0048】

一方、DTD+パターンツリー作成部51は、オペレータによって選択されたDTD+パターン情報Rを読み込んで(読込部に相当)、その内容を解析することによって、図7及び図12に示したようなDTD&パターンツリー55を、RAM3上に構築する。

【0049】

全体コントロール部52は、DTD+パターンツリー作成部51が作成したDTD+パターンツリー55における各要素のパターン記述情報を順次読み出して、読み出したパターン記述情報に沿った処理対象文書T中の領域の抽出を、パターン検索部53に対して依頼する。この際に、全体コントロール部52は、その要素について繰り返し有りとは指定されている場合には、パターン検索部53が検

索結果を報告できなくなるまで、その要素のパターン記述情報に沿った領域の抽出を、パターン検索部 5 3 に対して依頼する。そして、全体コントロール部 5 2 は、各要素毎になした抽出依頼に応じてパターン検索部 5 3 が抽出して来た処理対象文書 T 中の領域を、DTD+パターンツリー 5 5 における各要素の位置（即ち、「DTD+パターン情報」R 中の DTD）に従って、図 9 及び図 1 7 に示したような出力結果ツリー 5 6 として組み上げ、最後に、出力結果ツリー 5 6 における各要素に対応した領域の前後にその要素に対応したタグを付加することによって、図 1 0 に示したような構造化文書 O を出力する（構造化文書生成部に相当）。

【 0 0 5 0 】

パターン検索部 5 3 は、全体コントロール部 5 2 から抽出を依頼された要素の記述パターン情報における記述パターン（要素全体のパターン、若しくは、開始パターン及び終了パターン）の種類に対応したパターン検索部 5 3（パターン記述が文字列そのものである場合には文字列検索部 5 3 1，記述パターンが正規表現である場合には正規表現検索部 5 3 2，記述パターンが構文要素である場合には構文要素検索部 5 3 3）を呼び出して、その記述パターンに対応した文字列の検索を命じる。この際に、パターン検索部 5 3 は、抽出対象の要素の親ノードについて既に抽出した領域を、検索対象範囲として指定する。また、抽出対象の要素に順序性有りとは指定されている場合には、親ノードについて既に抽出した領域における兄ノードについて既に抽出した領域の後の部分を、検索対象範囲として指定する。また、抽出対象の要素に繰り返し有りの指定がなされている場合において、全体コントロール部 5 2 から同じ要素の抽出を 2 回目以降に依頼された場合には、親ノードについて既に抽出した領域におけるその要素について前回に抽出した領域の後の部分を、検索対象範囲として指定する。なお、パターン検索部 5 3 は、開始パターンと終了パターンとで記述パターンの種類が異なるのであれば、夫々の記述パターンに対応して文字列検索部 5 3 1 及び正規表現検索部 5 3 2 を呼び出して、夫々の記述パターンに対応した文字列の検索を命じる。

【 0 0 5 1 】

パターン検索部 5 3 は、文字列検索部 5 3 1，正規表現検索部 5 3 2 又は構文

要素検索部 5 3 3 が検索結果を報告して来ると（文字列検索部 5 3 1 及び正規表現検索部 5 3 2 に開始パターンと終了パターンとに対応した文字列の検索を命じた場合ににおいては、両検索部 5 3 1, 5 3 2 からの検索結果の報告が揃うと）、その検索結果を参照して、その要素に対応する領域（要素全体の記述パターンが指定されている場合には検索された文字列、開始パターン及び終了パターンが指定されている場合には、検索された文字列に挟まれた領域、但し、開始パターン又は終了パターンについて「内容にパターンを含む」と指定されている場合にはその記述パターンをも含む領域）を処理対象文書 T から抽出して、抽出した領域を全体コントロール部 5 2 に通知する（検索部に相当）。

【 0 0 5 2 】

文字列検索部 5 3 1 は、パターン検索部 5 3 から命じられた記述パターンと全く同じ文字列を検索し、正規表現検索部 5 3 2 は、パターン検索部 5 3 から命じられた記述パターンが示す正規表現に合致する文字列を検索し、構文要素検索部は、パターン検索部 5 3 から命じられた記述パターンと同じ構文要素（又は／及び、その前又は後に連続したコメント部）を検索し、パターン検索部 5 3 に報告する。

【 0 0 5 3 】

以上のような各モジュールから構成される構造文書化システム 5 は、入力装置 8 を介してオペレータによって入力される開始コマンドによって起動し、同じくオペレータによる入力によって処理対象文書 T 及び DTD + パターン情報 R が選択されることによって、図 4 に示す手順にて処理を開始する。

【 0 0 5 4 】

図 4 において、スタート後最初の S 0 0 1 では、DTD + パターンツリー作成部 5 1 が、選択された DTD + パターン情報 R を、ハードディスク 2 から RAM 3 上に読み込む。

【 0 0 5 5 】

次の S 0 0 2 では、DTD + パターンツリー作成部 5 1 が、S 0 0 1 にて読み込んだ DTD + パターン情報 R に基づいて、DTD + パターンツリー 5 5 を、RAM 3 上で構築する。

【0056】

次のS003では、全体コントロール部52が、選択された処理対象文書Tをハードディスク2からRAM3上に読み込む。

【0057】

次のS004では、全体コントロール部52が、S002にて作成されたDTD+パターンツリー55が、構文要素からなる記述パターンを含むか否かをチェックする。そして、DTD+パターンツリー55が構文要素からなる記述パターンを含まない場合には、全体コントロール部52は、S006において処理対象文書T自体を検索対象として決定した後に、処理をS007に進める。これに対して、DTD+パターンツリー55が構文要素からなる記述パターンを含む場合には、全体コントロール部52は、S005において、構文分解定義ファイル8を読み込んで、この構文分解定義ファイルBを参照することによって、処理対象文書Tに基づいて構文木・コメント木57を作成し、この構文木・コメント木57を検索対象として決定した後に、処理をS007に進める。

【0058】

S007では、全体コントロール部52が、DTD+パターンツリー55に従って出力結果ツリー56を作成する処理を、実行する。図5及び図6は、このS007にて実行される出力結果ツリー作成処理サブルーチンを示すフローチャートである。このサブルーチンに入って最初のS101では、全体コントロール部52が、DTD+パターンツリー55におけるルートノードに対応した領域が処理対象文書全体であると決定して、処理対象文書T全体をルートノードに対応した抽出結果とする出力結果ツリー56を生成する。

【0059】

次のS102では、全体コントロール部52は、DTD+パターンツリー55におけるルートノードの長子ノードを、処理対象ノードとする。次に、全体コントロール部52は、S103乃至S113のループ処理を実行する。このループ処理に入って最初のS103では、全体コントロール部52は、DTD+パターンツリー55における処理対象ノードから、その要素に指定された記述パターン情報を取り出す。

【 0 0 6 0 】

次の S 1 0 4 では、全体コントロール部 5 2 は、処理対象ノードの親ノードに対応した領域内（構文木・コメント木 5 7 にあっては親ノードの下位階層）を、S 1 0 3 にて取り出した記述パターンに合致する領域（要素全体の記述パターンが指定されている場合には検索された文字列、開始パターン及び終了パターンが指定されている場合には、検索された文字列に挟まれた領域、但し、開始パターン又は終了パターンについて「内容にパターンを含む」と指定されている場合にはその記述パターンをも含む領域）の検索対象範囲として、決定する。

【 0 0 6 1 】

次の S 1 0 5 では、パターン検索部 5 3 は、親ノードの領域内における検索開始位置を、処理対象ノードの特性（順序性の有無、兄ノードの有無、繰り返し指定があるノードに対する 2 回目以降の処理であるか否か）に従って、決定する。即ち、順序性が有り且つ兄ノードがある場合（但し、繰り返し指定があるが当該処理対象ノードに対する 2 回目以降の処理である場合を除く）には、パターン検索部 5 3 は、S 1 0 6 において、直前の兄ノードに対応した検索済み領域の後から検索することとする。また、繰り返し指定があるが当該処理対象ノードに対する 2 回目以降の処理である場合には、パターン検索部 5 3 は、S 1 0 7 において、当該処理対象ノードに関して前回の処理で検索した領域の後ろから検索することとする。また、繰り返し指定無し且つ順序性無しの場合や、その他の場合には、パターン検索部 5 3 は、S 1 0 8 において、親ノードの先頭から検索することとする。

【 0 0 6 2 】

何れの場合においても、次の S 1 0 9 において、パターン検索部 5 3 は、処理対象ノードの記述パターン情報における記述パターンの指定方法（要素全体の記述パターンを指定するか要素の開始パターン及び終了パターンを指定するか）及び表現方法（文字列のものを指定するか文字列の正規表現を指定するか）に従って、検索対象領域から、指定された記述パターンに合致した領域を検索する。この検索によって抽出された結果は、全体コントロール部 5 2 に通知される。

【 0 0 6 3 】

次の S 1 1 0 では、全体コントロール部 5 2 は、S 1 0 9 での検索の結果として処理対象ノードの記述パターンに合致した領域が検索対象領域から抽出されたか否かをチェックする。そして、合致した領域が抽出されがならば、全体コントロール部 5 2 は、S 1 1 1 において、抽出された領域の文字列を内容とするノードを、出力結果ツリー 5 6 における親ノードの下位階層に追加する。

【 0 0 6 4 】

次の S 1 1 2 では、全体コントロール部 5 2 は、処理対象ノードに子ノードがあるか否かをチェックする。そして、処理対象ノードに子ノードがあるならば、全体コントロール部 5 2 は、S 1 1 3 において、現在の処理対象ノードの長子ノードを、新たな処理対象ノードとして、処理を S 1 0 3 に戻す。

【 0 0 6 5 】

以上説明した S 1 0 3 乃至 S 1 1 3 のループ処理を繰り返した結果、S 1 0 9 での検索の結果として処理対象ノードの記述パターンに合致した領域が検索対象領域から抽出されなかったと S 1 1 0 にて判定した場合には、全体コントロール部 5 2 は、S 1 1 4 において、現在の処理対象ノードに対応した領域は無いものとして、空文字列を内容とするノードを出力結果ツリー 5 6 における親ノードの下位階層に追加する。この S 1 1 4 の完了後、全体コントロール部 5 2 は、処理を S 1 1 6 に進める。

【 0 0 6 6 】

また、上述した S 1 0 3 乃至 S 1 1 3 のループ処理を繰り返した結果、処理対象ノードに子ノードが無いと S 1 1 2 にて判定した場合（処理対象ノードがいわゆる葉ノードである場合）には、全体コントロール部 5 2 は、処理を S 1 1 5 に進める。

【 0 0 6 7 】

S 1 1 5 では、全体コントロール部 5 2 は、処理対象ノードに繰り返し指定があるか否かをチェックする。そして、繰り返し指定がある場合には、全体コントロール部 5 2 は、処理対象ノードをそのままにして、処理を S 1 0 3 に戻す。

【 0 0 6 8 】

これに対して、処理対象ノードに繰り返し指定がないと判定した場合には、全

体コントロール部 5 2 は、処理を S 1 1 6 に進める。

【 0 0 6 9 】

S 1 1 6 では、全体コントロール部 5 2 は、処理対象ノードに弟ノードがあるか否かをチェックする。そして、弟ノードがある場合には、全体コントロール部 5 2 は、S 1 1 7 において、次弟ノードを新たな処理対象ノードとして、処理を S 1 0 3 に戻す。

【 0 0 7 0 】

これに対して、処理対象ノードに弟ノードが無いと判定した場合には、全体コントロール部 5 2 は、S 1 1 8 にて現在の処理対象ノードの親ノードを仮の処理対象ノードとして、処理を S 1 1 9 に進める。S 1 1 9 では、全体コントロール部 5 2 は、仮の処理対象ノードがルートノードであるか否かをチェックする。そして、仮の処理対象ノードがルートノードではないならば、全体コントロール部 5 2 は、処理を S 1 1 5 に戻す。この場合、全体コントロール部 5 2 は、仮の処理対象ノードに繰り返し指定があるか否かをチェックして、繰り返し指定があれば当該仮の処理対象ノードを本来の処理対象ノードとして扱って、処理を S 1 0 3 に戻す。これに対して、仮の処理対象ノードに繰り返し指定がなければ、当該仮の処理対象ノードが弟ノードを有している否かをチェックする。そして、当該仮の処理対象ノードが弟ノードを有していれば、当該弟ノードを新たな処理対象ノードとし（S 1 1 7）、弟ノードを有していなければ、当該仮の処理対象ノードの更に親ノードを仮の処理対象ノードとする（S 1 1 8）。

【 0 0 7 1 】

以上説明した S 1 0 3 乃至 S 1 1 9 の処理を繰り返すことによって、DTD + パターンツリー 5 5 を構成する全てのノードに基づいた検索が行われる。そして、全ノードに基づいた検索が完了すると、S 1 1 9 にて仮の処理対象ノードがルートノードであるとの判定がなされ、この出力結果ツリー作成サブルーチンが終了し、処理を図 4 のメインルーチンに戻る。従って、この時点において、出力結果ツリー 5 6 が完成する。

【 0 0 7 2 】

処理が戻された図 4 のメインルーチンでは、処理は、S 0 0 7 から S 0 0 8 に

進む。このS008では、全体コントロール部52は、S007での処理の結果完成した出力結果ツリー56に基づいて、構造化文書Oを作成する。具体的には、全体コントロール部52は、子ノードを有していないノード（いわゆる葉ノード）に対応した領域の前後に、そのノード（要素）に対応付けられたタグを付加する。次に、兄弟ノード同士をひとまとめにして、その全体の前後に、これらノードに共通した親ノードに対応付けられたタグを付加する。このようにして、最下位階層ノードから上位ノードへと順次タグを付加して行き、最終的に、ルートノードに対応付けられたタグを付加することによって、構造化文書Oが完成する。全体コントロール部52は、このようにして完成した構造化文書Oを、ハードディスク2及びディスプレイ4に出力する。

【0073】

次のS009では、全体コントロール部52は、S001にて読み込んだDTD+パターン情報Rに基づいて処理すべき他の処理対象文書Tがオペレータによって選択されているか否かをチェックする。そして、他の処理対象文書Tがオペレータによって選択されていない場合には、全体コントロール部52は、処理をS003に戻す。

【0074】

これに対して、他の処理対象文書Tがオペレータによって選択されている場合には、全体コントロール部52は、S010において、現在参照しているDTD+パターン情報Rを変更する旨がオペレータによって入力されているか否かをチェックする。そして、DTD+パターン情報Rを変更する旨が入力されている場合には、全体コントロール部52は、処理をS001に戻す。これに対して、DTD+パターン情報Rを変更する旨が入力されていない場合には、構造化文書化システム5による処理が終了する。

（構造化文書化システムの動作例）

次に、以上の様な手順で処理を実行する構造化文書化システム5による具体的な動作例を、説明する。

【0075】

いま、図19に示すような内容を有するDTD+パターン情報Rがオペレータ

によって選択されるとともに、図21に示すような内容を有する処理対象文書Tがオペレータによって選択されたとする。すると、構造文書化システム5のDTD+パターンツリー作成部51は、このDTD+パターン情報Rの内容を解析して、図20に示すようなDTD+パターンツリー55を作成する(S001, S002)。

【0076】

全体コントロール部52は、このDTD+パターンツリー55を参照し、最初に、ルートノード「開発履歴」に対応する領域が、この処理対象文書Tの全体であると決定する(S101)。次に、全体コントロール部52は、ルートノードの子ノードを、順を追って処理対象ノードとしていく(S102, S103～S113)。

【0077】

先ず、全体コントロール部52は、ルートノードの長子ノード「初版情報」を、処理対象ノードとする(S102)。そして、全体コントロール部52は、DTD+パターンツリー55から、このノード「初版情報」についての記述パターン情報を参照して(S103)、親ノード「開発履歴」に対応する領域(処理対象文書Tの全体)を検索対象とする(S104)。そして、この記述パターン情報には繰り返し及び順序性が共に無しと指定されているので、パターン検索部53は、親ノード「開発履歴」に対応する領域の先頭から、検索を開始する(S108, S109)。この検索において、この記述パターン情報にはパターン指定方法が開始及び終了と指定され、開始パターンが文字列そのものによる“初版作成者”、終了パターンが正規表現による“《行末》”と指定されているので、

“藤川 泰之 : 1999.01.01”

の部分が、記述パターン情報に合致した領域として検出される。従って、この領域が、ノード「初版情報」に対応した領域として抽出され、出力結果ツリー56に追加される(S111)。

【0078】

次に、全体コントロール部52は、その長子ノード「作成者」を、新たな処理対象ノードとする(S112, S113)。そして、全体コントロール部52は

、DTD+パターンツリー55から、このノード「作成者」についての記述パターン情報を参照して(S103)、親ノード「初版情報」に対応する領域

“藤川 泰之 : 1999.01.01”

を検索対象とする(S104)。そして、この記述パターン情報には繰り返し及び順序性が共に無しとして指定されているので、パターン検索部53は、親ノード「初版情報」に対応する領域の先頭から、検索を開始する(S108, S109)。この検索において、この記述パターン情報にはパターン指定方法が開始及び終了と指定され、開始パターンが正規表現による“《領域の先頭》”、終了パターンが文字列そのものによる“:”と指定されているので、

“藤川 泰之”

の部分が、記述パターン情報に合致した領域として検出される。従って、この領域が、ノード「作成者」に対応した領域として抽出され、出力結果ツリー56に追加される(S111)。

【0079】

このノード「作成者」には子ノードがなく(S112)、その記述パターン情報には繰り返しが無しと指定されているので(S115)、全体コントロール部52は、その次弟ノード「作成日時」を、新たな処理対象ノードとする(S116, S117)。そして、全体コントロール部52は、DTD+パターンツリー55から、このノード「作成日時」についての記述パターン情報を参照して(S103)、親ノード「初版情報」に対応する領域

“藤川 泰之 : 1999.01.01”

を検索対象とする(S104)。そして、この記述パターン情報には繰り返しが無しと指定されているが順序性が有りとして指定されているので、パターン検索部53は、兄ノード「作成者」の後から、検索を開始する(S106, S109)。この検索において、この記述パターン情報にはパターン指定方法が開始及び終了と指定され、開始パターンが文字列そのものによる“:”、終了パターンが正規表現による“《行末》”、と指定されているので、

“1999.01.01”

の部分が、記述パターン情報に合致した領域として検出される。従って、この領

域が、ノード「作成日時」に対応した領域として抽出され、出力結果ツリー 5 6 に追加される (S 1 1 1)。

【0080】

このノード「作成者」には子ノードがなく (S 1 1 2)、その記述パターン情報には繰り返しが無しと指定されており (S 1 1 5)、しかも弟ノードがないので (S 1 1 6)、全体コントロール部 5 2 は、親ノード「初版情報」の次弟ノード「更新履歴」を、新たな処理対象ノードとする (S 1 1 8, S 1 1 9, S 1 1 5 ~ S 1 1 7)。そして、全体コントロール部 5 2 は、DTD+パターンツリー 5 5 から、このノード「更新履歴」についての記述パターン情報を参照して (S 1 0 3)、親ノード「開発履歴」に対応する領域 (処理対象文書 T の全体) を検索対象とする (S 1 0 4)。そして、この記述パターン情報には順序性有りとは指定されているので、パターン検索部 5 3 は、兄ノード「初版情報」の後から、検索を開始する (S 1 0 6, S 1 0 9)。この検索において、この記述パターン情報にはパターン指定方法が開始及び終了と指定され、開始パターンが文字列そのものによる“更新履歴”、終了パターンが正規表現による“《行末》”、と指定されているので、

“1999.12.16 / 第1.1版”

の部分が、記述パターン情報に合致した領域として最初に検出される。従って、この領域が、ノード「作成日時」に対応した領域として抽出され、出力結果ツリー 5 6 に追加される (S 1 1 1)。

【0081】

次に、全体コントロール部 5 2 は、その長子ノード「更新日付」を、新たな処理対象ノードとする (S 1 1 2, S 1 1 3)。そして、全体コントロール部 5 2 は、DTD+パターンツリー 5 5 から、このノード「更新日付」についての記述パターン情報を参照して (S 1 0 3)、親ノード「更新履歴」に対応して抽出された領域

“1999.12.16 / 第1.1版”

を検索対象とする (S 1 0 4)。そして、この記述パターン情報には繰り返し及び順序性が共に無しとして指定されているので、パターン検索部 5 3 は、親ノード

ド「更新履歴」に対応する領域の先頭から、検索を開始する（S 1 0 8, S 1 0 9）。この検索において、この記述パターン情報にはパターン指定方法が開始及び終了と指定され、開始パターンが正規表現による“《領域の先頭》”、終了パターンが文字列そのものによる“／”と指定されているので、

“1999.12.16”

の部分が、記述パターン情報に合致した領域として検出される。従って、この領域が、ノード「更新日付」に対応した領域として抽出され、出力結果ツリー 5 6 に追加される（S 1 1 1）。

【 0 0 8 2 】

このノード「更新日付」には子ノードがなく（S 1 1 2）、その記述パターン情報には繰り返しが無しと指定されているので（S 1 1 5）、全体コントロール部 5 2 は、その次弟ノード「版数」を、新たな処理対象ノードとする（S 1 1 6, S 1 1 7）。そして、全体コントロール部 5 2 は、DTD+パターンツリー 5 5 から、このノード「版数」についての記述パターン情報を参照して（S 1 0 3）、親ノード「更新情報」に対応して抽出された領域

“1999.12.16 / 第1.1版”

を検索対象とする（S 1 0 4）。そして、この記述パターン情報には繰り返しが無しと指定されているが順序性が有りとは指定されているので、パターン検索部 5 3 は、兄ノード「更新日付」の後から、検索を開始する（S 1 0 6, S 1 0 9）。この検索において、この記述パターン情報にはパターン指定方法が開始及び終了と指定され、開始パターンが文字列そのものによる“／”、終了パターンが正規表現による“《行末》”、と指定されているので、

“第1.1版”

の部分が、記述パターン情報に合致した領域として検出される。従って、この領域が、ノード「版数」に対応した領域として抽出され、出力結果ツリー 5 6 に追加される（S 1 1 1）。

【 0 0 8 3 】

このノード「版数」には子ノードがなく（S 1 1 2）、その記述パターン情報には繰り返しが無しと指定されており（S 1 1 5）、しかも弟ノードがないので

(S116)、全体コントロール部52は、親ノード「更新履歴」を仮の処理対象ノードとする(S118)。当該仮の処理対象ノード「更新履歴」の記述パターン情報には繰り返し有りとして指定されているので(S115)、全体コントロール部52は、このノード「更新履歴」による領域抽出を繰り返す。この場合、処理が2回目となるので、全体コントロール部52は、親ノード「開発履歴」に対応する領域(処理対象文書Tの全体)における前回の処理で抽出された領域

“1999.12.16 / 第1.1版”

の後から、検索を開始する(S107, S109)。この検索においては、

“2000.02.14 / 第1.2版”

の部分が、記述パターン情報に合致した領域として最初に検出される。また、その後で行われるノード「更新日付」、ノード「版数」についての検索では、

“2000.02.14”

“第1.2版”

の部分が、夫々検出される。

【0084】

その後においては、全体コントロール部52は、再度、ノード「更新履歴」についての検索を試みるが、もはや記述パターンに合致する領域は検出されず(S110)、このノード「更新履歴」には弟ノードもないので、全体コントロール部52は、一旦、親ノード「開発履歴」を仮の処理対象ノードとし(S118)、当該処理対象ノード「開発履歴」がルートノードであるが故に(S119)、検索及び出力結果ツリー55の作成を終了する。この時点におけるDTD+パターンツリー55は、図22に示す通りになる。

【0085】

全体コントロール部52は、このDTD+パターンツリー55に基づいて、その各ノードに付された文字列に対して、タグを付加することによって、図23に示すような構造化文書を作成して出力する(S008)。

(構造化文書化システムの詳細構成及び処理内容)

次に、上述したDTD+パターン作成支援システム7による処理内容を、詳細に説明する。図27は、このDTD+パターン作成支援システム7の処理内容(

D T D + パターン作成支援システム 7 を構築するプログラムによる C P U 1 の処理内容) を示すフローチャートである。

【0086】

この D T D + パターン作成支援システム 7 は、入力装置 8 を介してオペレータによって入力される開始コマンドによって起動し、図 24 に示すような選択画面をディスプレイ 4 上に表示するとともに、この選択画面における各典型パターン選択ボタン 741 に、対応する典型パターン定義情報 S を関連付ける。そして、入力装置 8 を介してオペレータによる入力によって処理対象文書 T のサンプルが選択されると、D T D + パターン作成支援システム 7 は、この処理対象文書 T のサンプルをハードディスク 2 から R A M 3 上に読み込み、選択画面におけるサンプルプレビューボックス 72 にそのテキスト内容を表示する。そして、オペレータが、このサンプルプレビューボックス 72 内に表示されているテキストにおける何れかの行がドラッグすることによって選択した後で、その行のパターンに最も似ている典型パターンを探し出して、その典型パターンに対応した典型パターン選択ボタン 741 をクリックすると、D T D + パターン作成支援システム 7 は、図 27 の処理を開始する。

【0087】

図 27 の処理では、D T D + パターン作成支援システム 7 は、スタート後最初の S 201 において、オペレータが選択した行を、R A M 3 上の作業領域に読み込む。

【0088】

次の S 202 では、D T D + パターン作成支援システム 7 は、オペレータがクリックした典型パターン選択ボタン 741 に関連付けられた典型パターン定義情報 S を、R A M 3 上の作業領域に読み込む。そして、読み込んだ典型パターン定義情報 S の構造特定部 S 1 に記載されている典型パターンの概略構造を、分解する。具体的には、典型パターンの概略構造における各エレメント（隅付括弧で囲まれた部分を、他の部分から抽出する。

【0089】

次の S 203 では、D T D + パターン作成支援システム 7 は、S 202 にて分

解された各エレメント（隅付括弧で囲まれた部分）を先頭から一つずつ検索対象として特定し、各検索対象エレメントに関して文字種特定部 S 2 にて特定されている正規表現パターンに合致した部分を、S 2 0 1 にて R A M 3 上の作業領域に読み込まれたテキストから探し出す。このとき、D T D + パターン作成支援システム 7 は、最初のエレメントを検索対象とする場合には、S 2 0 1 にて R A M 3 上の作業領域に読み込まれたテキストの先頭から検索を行い、それ以降のエレメントを検索対象とする場合には、直前のエレメントに関して探し出された部分の後から検索を行う。

【0090】

次の S 2 0 4 では、D T D + パターン作成支援システム 7 は、図 2 8 に示すようなダイアログ 7 0 0 を、選択画面に重ねて表示する。このダイアログ 7 0 0 は、典型パターン定義情報 S 毎に作成される。図 2 8 に示す例は、図 2 5 に示される典型パターン定義情報 S に関連付けて作成されたものである。要素名テキストボックス 7 0 1、タイトル文字列テキストボックス 7 0 2、タイトル文字列リストボックス 7 0 3、区切り文字列テキストボックス 7 0 4、区切り文字列リストボックス 7 0 5、追加ボタン 7 0 6 が、含まれている。D T D + パターン作成支援システム 7 は、S 2 0 3 にて各エレメントに関して探し出された部分を、対応するテキストボックス 7 0 2、7 0 4 に表示する。

【0091】

図 2 8 は、サンプルプレビューボックス 7 2 内に表示されたテキスト中の

“会社名：富士通株式会社”

の行が選択された後に、図 2 5 に示す典型パターン情報 S に関連付けられた典型パターン選択ボタン 7 4 1 がクリックされた場合を示している。エレメント [タイトルパターン] に関して探し出された部分 “会社名” がタイトル文字列テキストボックス 7 0 2 に表示されているとともに、エレメント [区切りパターン] に関して探し出された部分 “：” が区切り文字列テキストボックス 7 0 4 に表示されている。

【0092】

なお、タイトル文字列リストボックス 7 0 3 には、オペレータが、タイトル文

字列テキストボックス 7 0 2 内に表示されている文字列と置換可能な文字列を入力し得る。同様に、区切り文字列リストボックス 7 0 5 には、オペレータが、区切り文字列テキストボックス 7 0 4 内に表示されている文字列と置換可能な文字列を入力し得る。また、要素名テキストボックス 7 0 1 には、オペレータが、作成しようとする記述パターンが指定される要素の要素名を入力し得る。そして、追加ボタン 7 0 6 をオペレータがクリックすると、DTD+パターン作成支援システム 7 は、処理を S 2 0 5 に進める。

【 0 0 9 3 】

S 2 0 5 では、DTD+パターン作成支援システム 7 は、典型パターン情報 S 中の雛形部 S 3 における各記述パターン情報の雛形の [] 内に、ダイアログ 7 0 0 の各欄に表示された文字列を、雛形部 S 3 内で指定された表現（文字種特定部 S 2 で特定された表現よりは具体的な表現）に変換してから代入する。図 2 5 及び図 2 6 の例では、タイトル文字列テキストボックス 7 0 2 内に表示されている文字列が正規表現に変換されて「タイトル文字列 1」に代入され、タイトル文字列リストボックス 7 0 3 内に表示されている文字列が正規表現に変換されて「タイトル文字列 2」に代入され、区切り文字列テキストボックス 7 0 4 内に表示されている文字列が正規表現に変換されて「区切り文字列 1」に代入され、区切り文字列リストボックス 7 0 5 内に表示されている文字列が「区切り文字列 2」に代入される。これにより、雛形部 S 3 内の雛形が、要素名テキストボックス 7 0 1 内に表示されている要素名を有する要素に対して指定される記述パターン情報となって、DTD+パターン情報 R に追加される。図 2 9 は、図 2 8 に示す状態で追加ボタン 7 0 6 がクリックされた場合に作成される記述パターン情報を示す。なお、上述したように、この時点において、ツリーリストボックス 7 3 内には、図 3 0 に示すように、ルートノード「設計書」の子ノードとして、要素名「会社名」が表示される。

【 0 0 9 4 】

以後、オペレータが、サンプルプレビューボックス 7 3 内に表示されているテキスト中の任意の行を選択して、何れかの典型パターン選択ボタン 7 4 1 をクリックする毎に、新たな子ノード（若しくは、子ノード及び孫ノード）についての

記述パターン情報が作成されて、DTD+パターン情報Rに追加される。

【0095】

例えば、ルートノード「設計書」の子ノードとして要素「会社情報」が追加された状態において、サンプルプレビューボックス72内に表示されたテキスト中の

“ファイル名<日本語名>ファイル長

KOKYAKU-MASTER<顧客マスタ>200”

の行が選択された後に、図26に示す典型パターン情報Sに関連付けられた典型パターン選択ボタン741がクリックされた場合のダイアログ700'を、図31に示す。このダイアログ700'には、タイトル文字列テキストボックス702が5個含まれ、区切り文字列テキストボックス704が4個含まれている。また、追加ボタン706の代わりに、OKボタン707が含まれている。

【0096】

このOKボタン707がクリックされると、図26に示す典型パターン情報S中の各雛形部S3中の[]に、ダイアログ700'の各欄に表示された文字列に基づいて変換された文字列が、代入される。その結果として、ツリーリストボックス73内には、図32に示すように、ルートノード「設計書」の子ノード及び孫ノードとして、要素名「ファイル属性」等が表示される。

【0097】

【発明の効果】

以上説明したように、文書構造の各要素の抽出条件の内容を任意に設定しておき、テキスト形式で記述された処理対象電子文書にこの抽出条件を当てはめることによって、文書構造の各要素に該当する領域を抽出することができるので、抽出した各領域にその要素に対応したタグを付加することによって、自動的に構造化文書を生成することができる。

【図面の簡単な説明】

【図1】 本発明の実施の形態である構造文書化システムの概念をDTD+パターン編集システム及びDTD+パターン作成支援システムの概念とともに示す概念図

【図 2】 構造文書化システム等が実現されるコンピュータの構成を示すブロック図

【図 3】 構造文書化システムの詳細なモジュール構成を示すプログラム構成部

【図 4】 構造文書化システムによる処理手順を示すフローチャート

【図 5】 図 4 の S 0 0 7 にて実行される出力結果ツリー作成サブルーチンを示すフローチャート

【図 6】 図 4 の S 0 0 7 にて実行される出力結果ツリー作成サブルーチンを示すフローチャート

【図 7】 DTD+パターンツリーの構成例を示す図

【図 8】 処理対象文書のテキスト例を示す図

【図 9】 出力結果ツリーの構成例を示す図

【図 1 0】 構造化文書例を示す図

【図 1 1】 正規化表現のルールを示す表

【図 1 2】 DTD+パターンツリーの構成例を示す図

【図 1 3】 処理対象文書のテキスト例を示す図

【図 1 4】 BNF の定義の一部を示す表

【図 1 5】 構文要素の範囲を示す図

【図 1 6】 構文木・コメント木の構成例を示す図

【図 1 7】 出力結果ツリーの構成例を示す図

【図 1 8】 DTD+パターン編集システムによる編集画面例を示す図

【図 1 9】 DTD+パターン情報のテキスト例を示す図

【図 2 0】 DTD+パターンツリーの構成例を示す図

【図 2 1】 処理対象文書のテキスト例を示す図

【図 2 2】 出力結果ツリーの構成例を示す図

【図 2 3】 構造化文書例を示す図

【図 2 4】 DTD+パターン作成支援システムによる選択画面例を示す図

【図 2 5】 典型パターン定義情報のテキスト例を示す図

【図 2 6】 典型パターン定義情報のテキスト例を示す図

【図 2 7】 DTD+パターン作成支援システムによる処理手順を示すフロー
チャート

【図 2 8】 DTD+パターン作成支援システムによる選択画面例を示す図

【図 2 9】 DTD+パターン作成支援システムによって作成される記述パタ
ーン例を示す図

【図 3 0】 DTD+パターン作成支援システムによる選択画面例を示す図

【図 3 1】 DTD+パターン作成支援システムによる選択画面例を示す図

【図 3 2】 DTD+パターン作成支援システムによる選択画面例を示す図

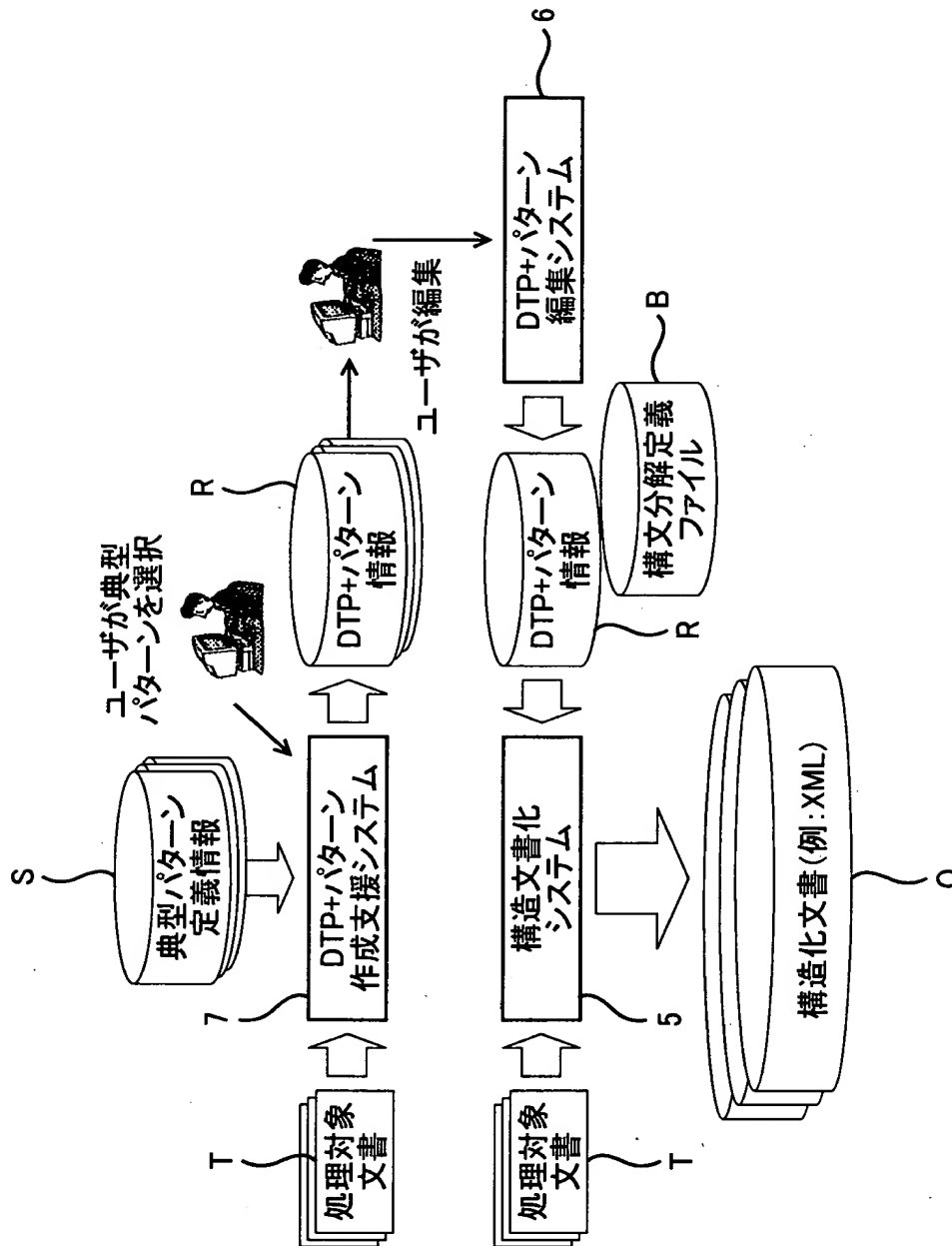
【図 3 3】 従来 of 構造化文書の説明図

【符号 of 説明】

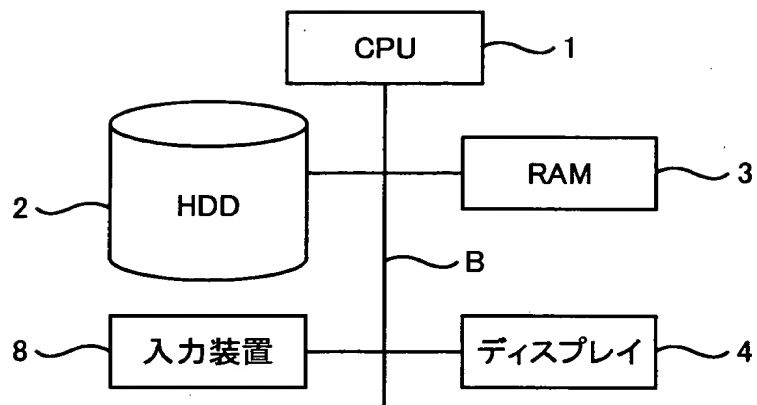
- | | |
|-----|------------|
| 1 | CPU |
| 2 | ハードディスク |
| 3 | RAM |
| 5 | 構造化文書化システム |
| 8 | 入力装置 |
| 5 2 | 全体コントロール部 |
| 5 3 | 検索部 |
| 5 4 | 構文木分解部 |
| R | DTD+パターン情報 |
| T | 処理対象文書 |
| O | 構造化文書 |

【書類名】 図面

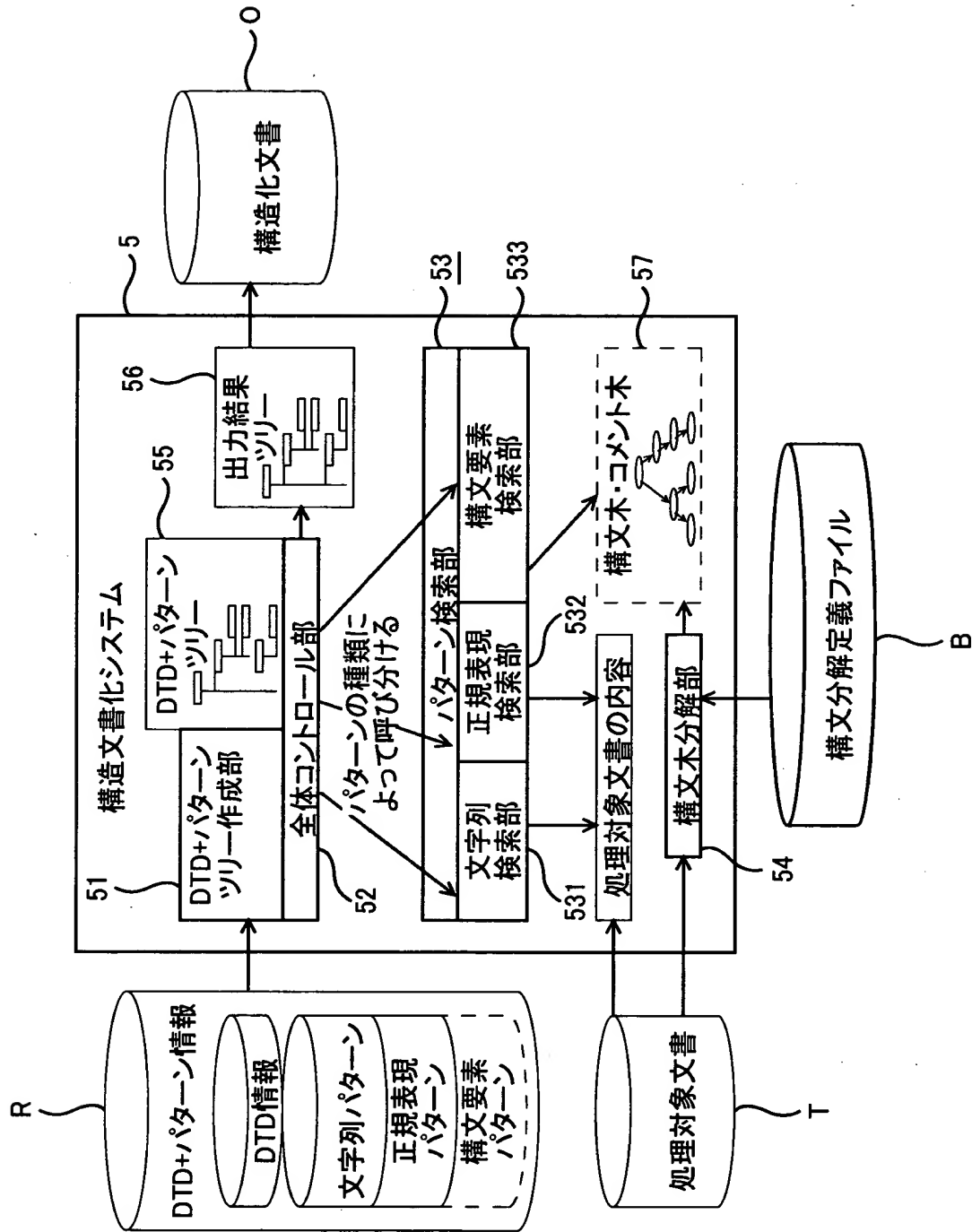
【図 1】



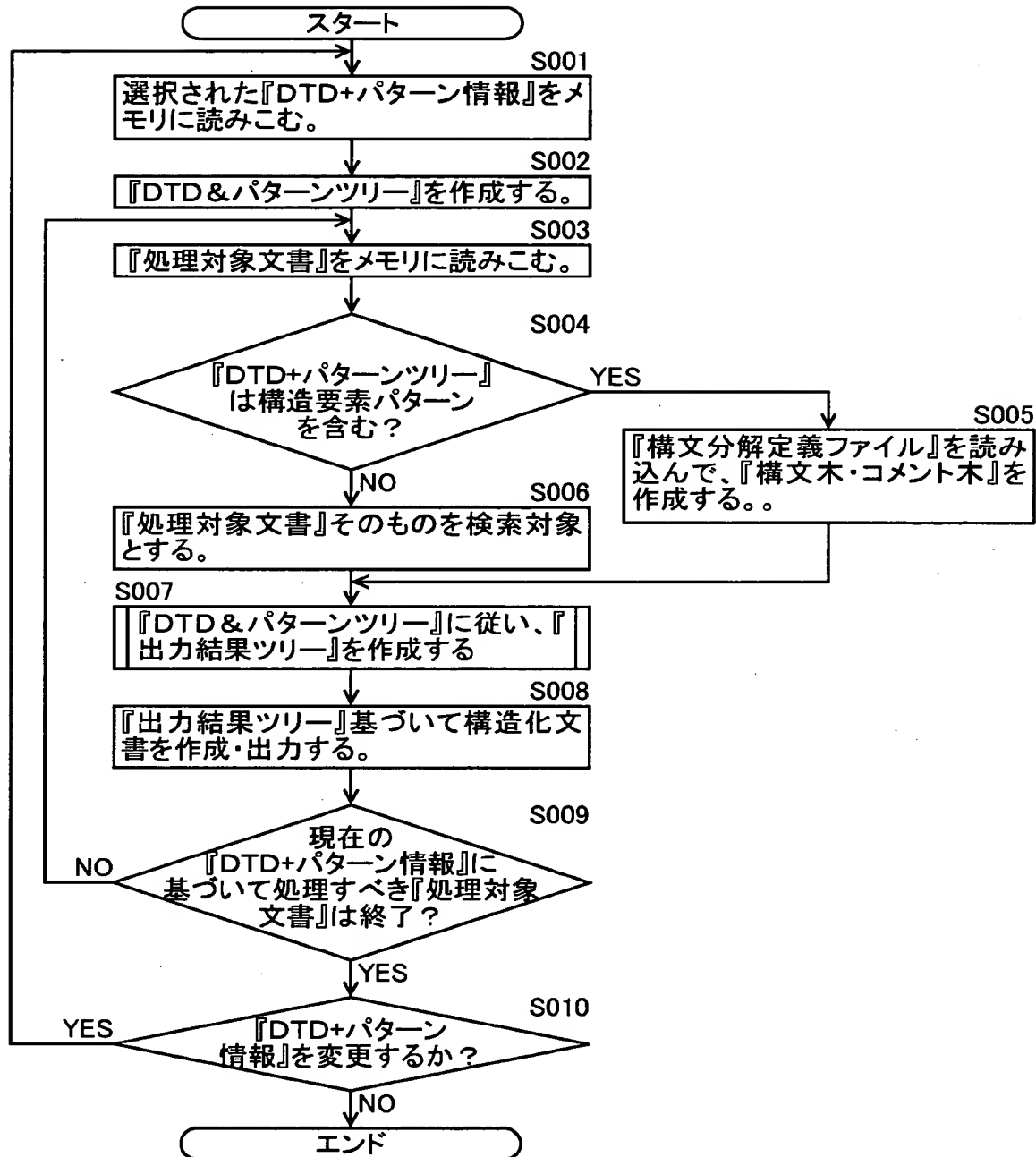
【図 2】



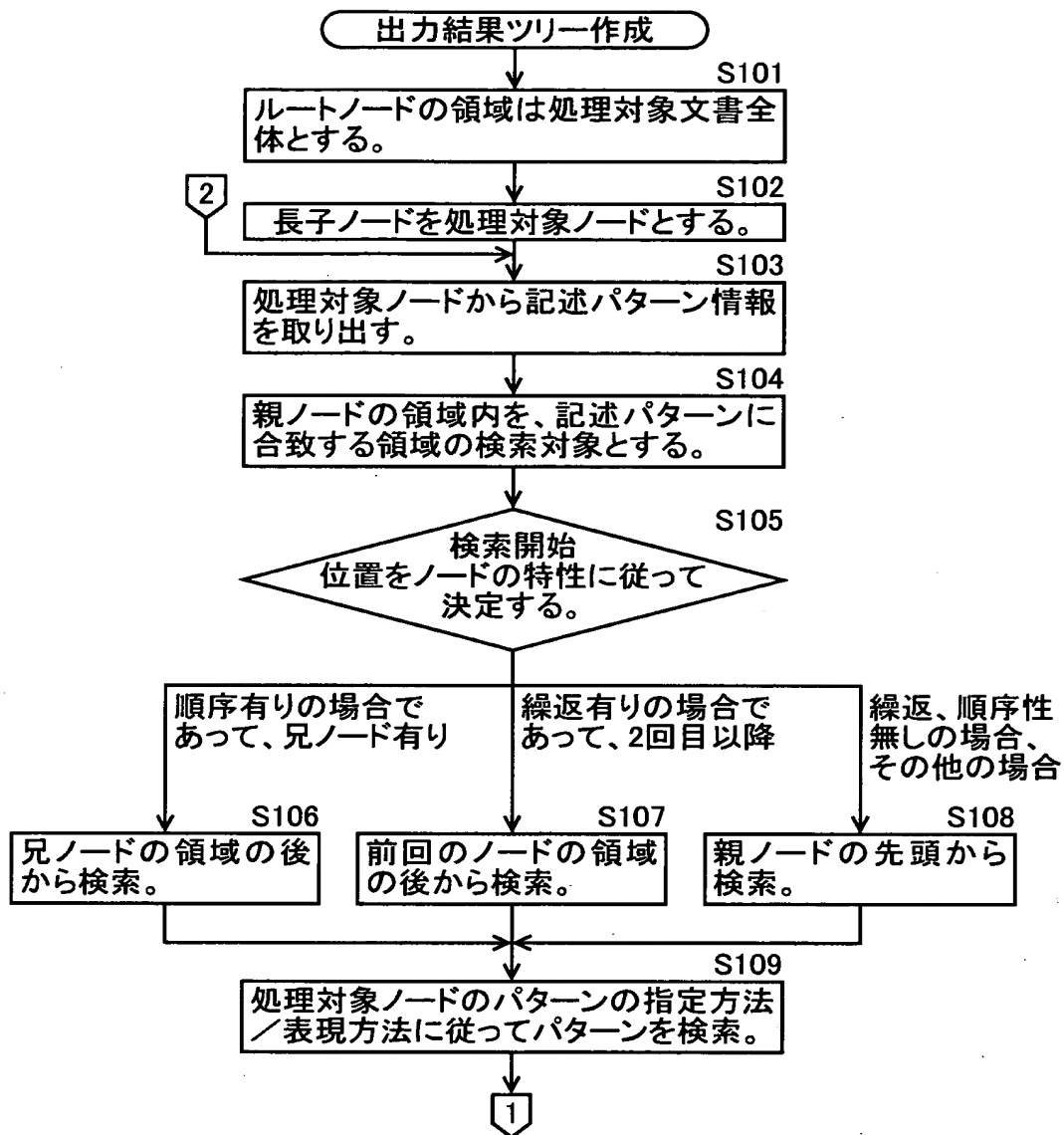
【図 3】



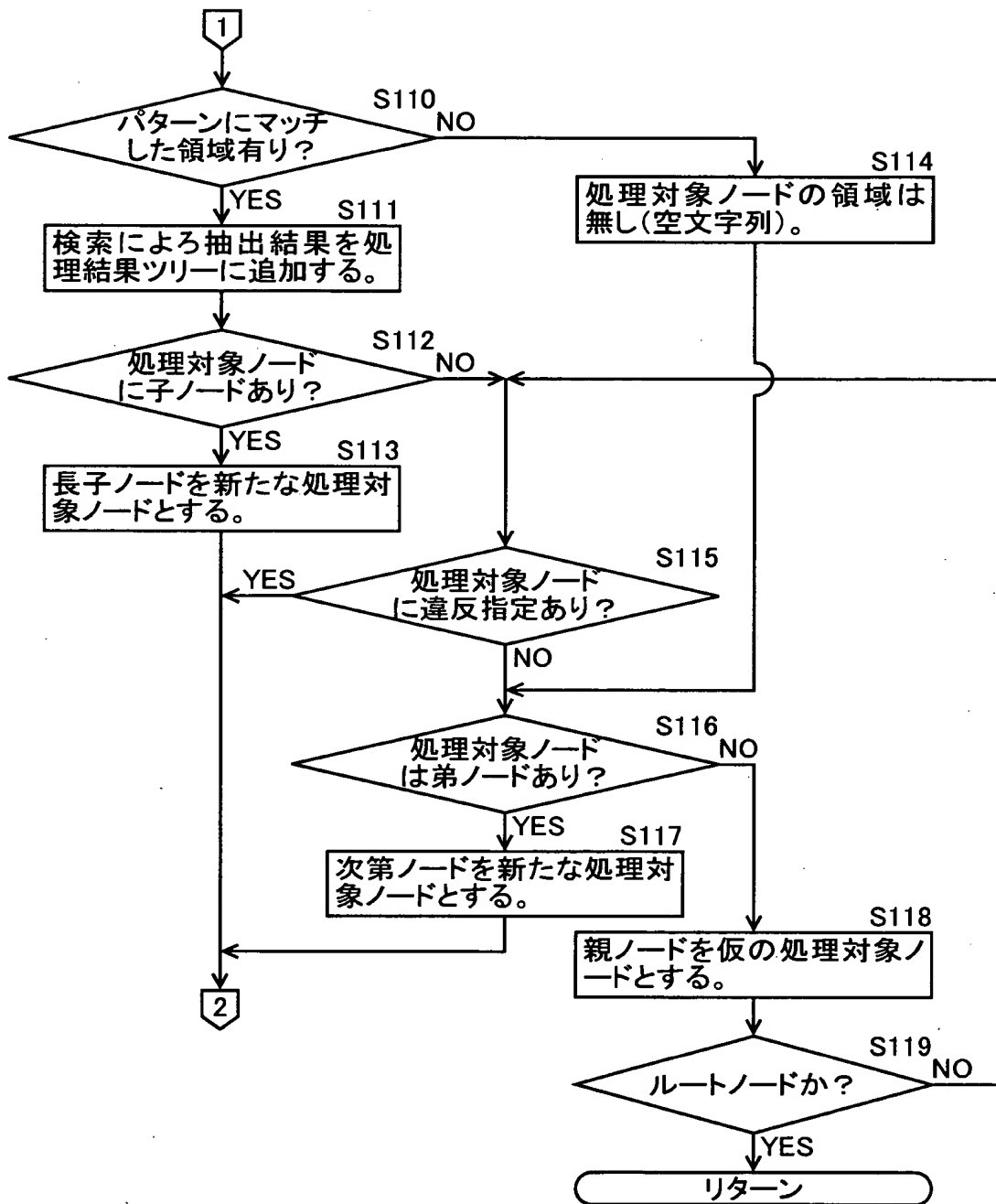
【図 4】



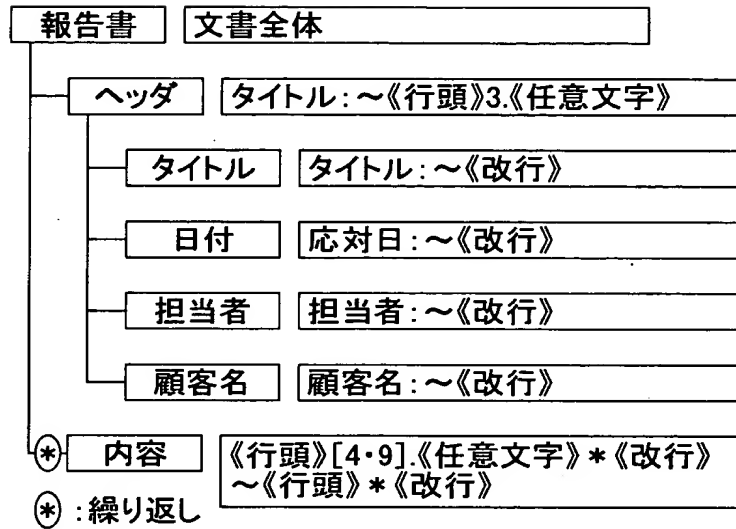
【図 5】



【図 6】



【図 7】



【図 8】

〇〇部長殿

タイトル: 商談対応報告書
担当者: 藤川 泰之

1. 顧客名 山田証券

2. 応対日: 1997.02.17

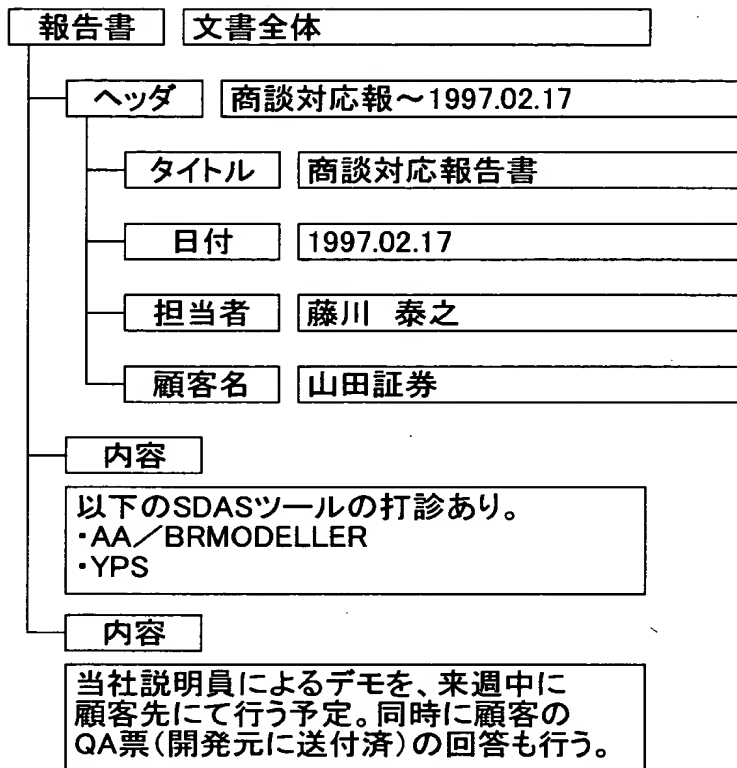
3. 概要
上記顧客より当社ツール使用の打診。

4. 概要
以下のSDASツールの打診あり。
・AA/BRMODELLER
・YPS

5. 今後の予定
当社説明員によるデモを、来週中に
顧客先にて行う予定。同時に顧客の
QA票(開発元に送付済)の回答も行う。

以上

【図 9】



【図 10】

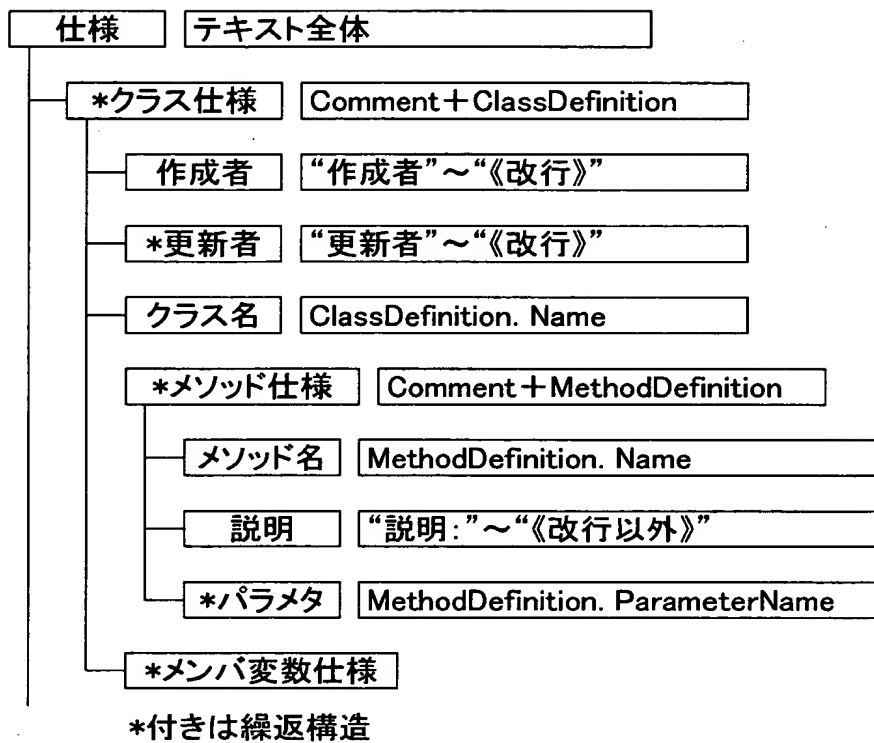
```

<報告書>
<ヘッダ>
<タイトル>商談対応報告書</タイトル>
<日付>1997.02.17</日付>
<担当者>藤川 泰之</担当者>
<顧客名>山田証券</顧客名>
</ヘッダ>
<内容>
以下のSDASツールの打診あり。
・AA/BRMODELLER
・YPS
</内容>
<内容>
当社説明員によるデモを、来週中に
顧客先にて行う予定。同時に顧客の
QA票(開発元に送付済)の回答も行う。
</内容>
<報告書>
  
```

【図 1 1】

正規表現	意味
.	改行以外の任意の1文字(本明細では可読性のため《任意文字》又は《改行以外》と表記している場合がある)
*	0個以上の繰り返し
+	1個以上の繰り返し
	OR
[]	かっこ内の任意の1文字
[^]	かっこ内の文字群以外の1文字
^	行頭(本明細では可読性のため《行頭》と表記している場合がある)
\$	行末(本明細では可読性のため《行末》と表記している場合がある)
()	優先順位
¥t	タブ(本明細では可読性のため《タブ》と表記している場合がある)
¥n	改行(本明細では可読性のため《改行》と表記している場合がある)
	上記記号を直接指定したい場合は、¥記号を前に置く

【図 1 2】



【図 1 3】

```

/** COPYRIGHT Fjitsu Ltd
 * 作成者 藤川 泰之(富士通)
 * 更新者 原田 義之(富士通)
 * 更新者 和田 憲明(富士通)
 */
public class 顧客 {
    /*
    *説明: 資本金から信用度を割出す。
    */
    public String 信用ランク(
        int 現在借入金,
        long 公定歩合)
    {
        :
        :
    }
    //説明: 資本金。
    public Static int 資本金;
}
  
```

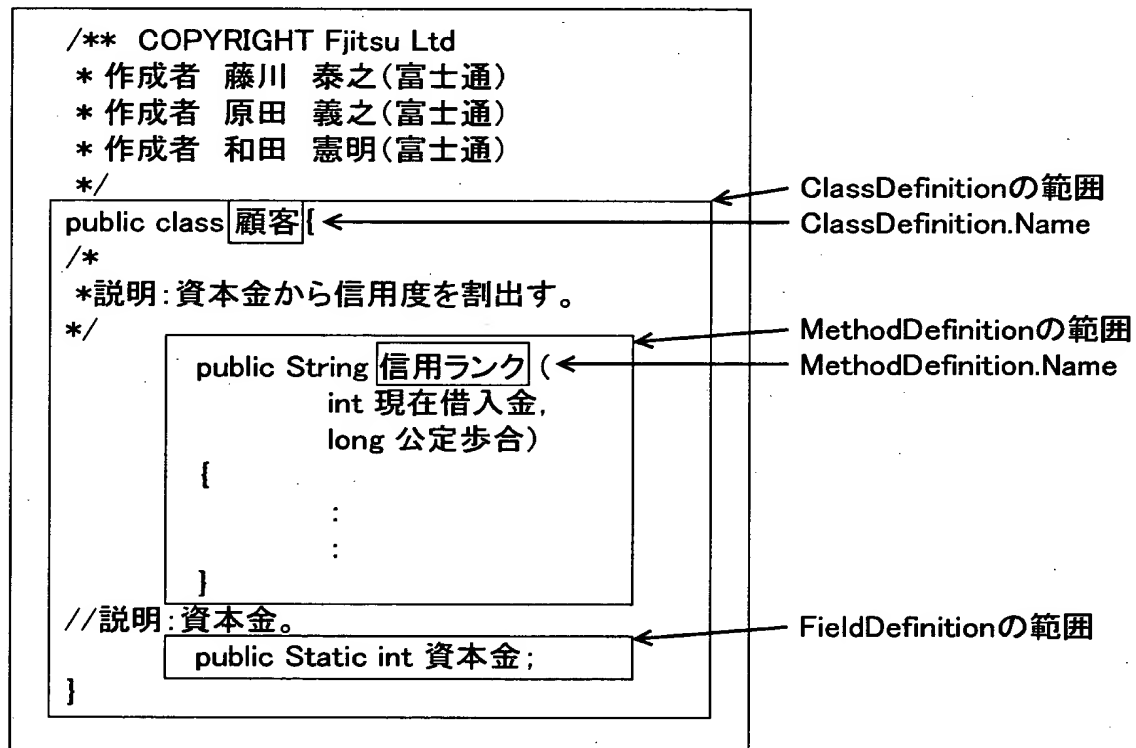
【図 1 4】

```

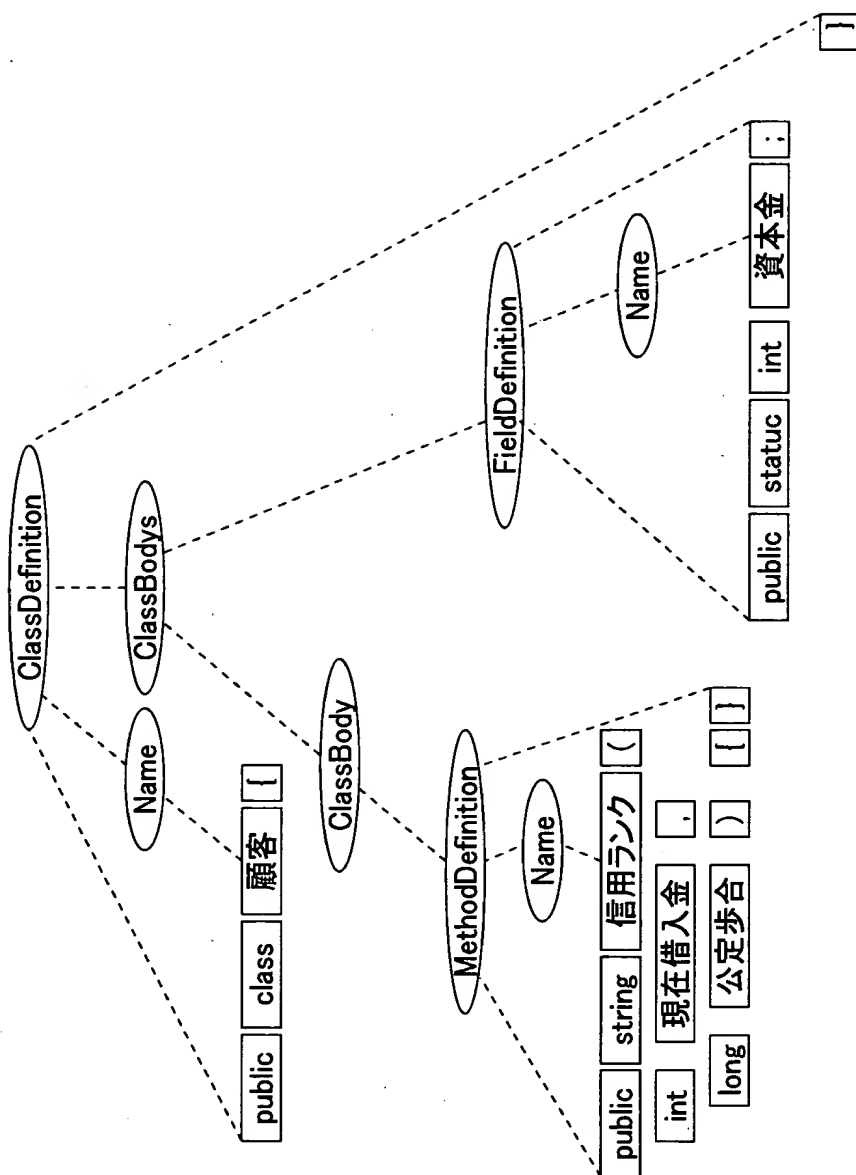
ClassDefinition: "public" "class" Name " { " ClassBodys " } ";
ClassBodys: ClassBody ";";
| ClassBodys ClassBodys ";";
ClassBody: FieldDefinition
| MethodDefinition;
MethodDefinition: "public" Type Name '(' FormalParameterList '(' '{' Block '}' ;
FormalParameterList: FormalParameter
| FormalParameterList ',' FormalParameter;
FormalParameter: TypeParameterName;
:
:
:

```

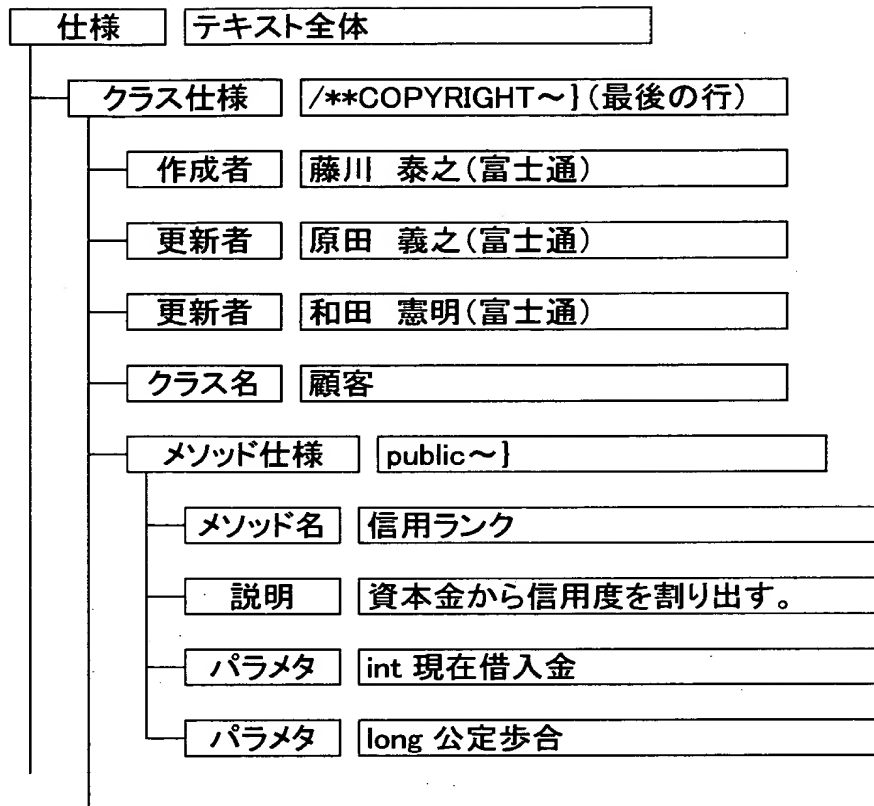
【図 1 5】



【図 16】



【図 17】



【図 18】

61

DTDツリー構造

622

要素名

タイトル

621

6210

6220

623

624

6242

6244

625

6254

626

627

63

64

65

66

67

68

削除

キャンセル

更新終了

内容反映

子として追加

単として追加

開始と終了

○ 開始と終了

○ パターンそのもの

パターンの意味

○ 繰り返しあり

前後空白除去

行間文字削除

パターン開始パターン

パターンタイプ

○ 文字列

○ 正規表現

○ 構文要素名

コメント処理

○ 前方コメント含む

○ 後方コメント含む

構文要素名参照

内容にパターンを含む

Title

終了パターン

パターンタイプ

○ 文字列

○ 正規表現

○ 構文要素名

内容にパターンを含む

Kn

構文要素名参照

制に対する範囲制限(親が構文要素の場合のみ)

○ 無し

○ 前方コメント

○ 構文要素

○ 後方コメント

出カタブ名

タイトル

要素名

タイトル

OSMインクフェース名

文字インクフェース名

パインクフェース名

顔クラス名(実型)

顔クラス名(カラ)

顔別

白インクフェース

顔南

顔青

白【参照定義】

インポートクラス名-1

インポートクラス名-2

インポートクラス名-3

白【プロパティ定義】

白【プロパティ名G

文字プロパティ名

型

初期値

プロパティの説明

プロパティの備考

【パング-算定補助】

日 マグダ名G

文字パング名

パング名

パングの説明

パングの備考

戻り値の型(プリミティブ)

戻り値の型(リファレンス)

パラメタ:型

【図19】

〔開発履歴〕

ルート: YES

子ノード=初版情報、更新履歴

〔初版情報〕

子ノード=作成者、作成日付

パターン指定方法: 開始／終了

開始パターン: 文字列: 「初版作成者」

終了パターン: 正規表現: 《行末》

繰返: 無し

順序性: 無し

〔作成者〕

パターン指定方法: 開始／終了

開始パターン: 正規表現: 《領域の先頭》

終了パターン: 文字列: 「:」

繰返: 無し

順序性: 無し

〔作成日付〕

パターン指定方法: 開始／終了

開始パターン: 文字列: 「:」

終了パターン: 正規表現: 《行末》

繰返: 無し

順序性: 有り

〔更新履歴〕

子ノード=更新日付、版数

パターン指定方法: 開始／終了

開始パターン: 文字列: 「更新履歴」

終了パターン: 正規表現: 《行末》

繰返: 有り

順序性: 有り

〔更新日付〕

パターン指定方法: 開始／終了

開始パターン: 正規表現: 《領域の先頭》

終了パターン: 文字列: 「/」

繰返: 無し

順序性: 無し

〔版数〕

パターン指定方法: 開始／終了

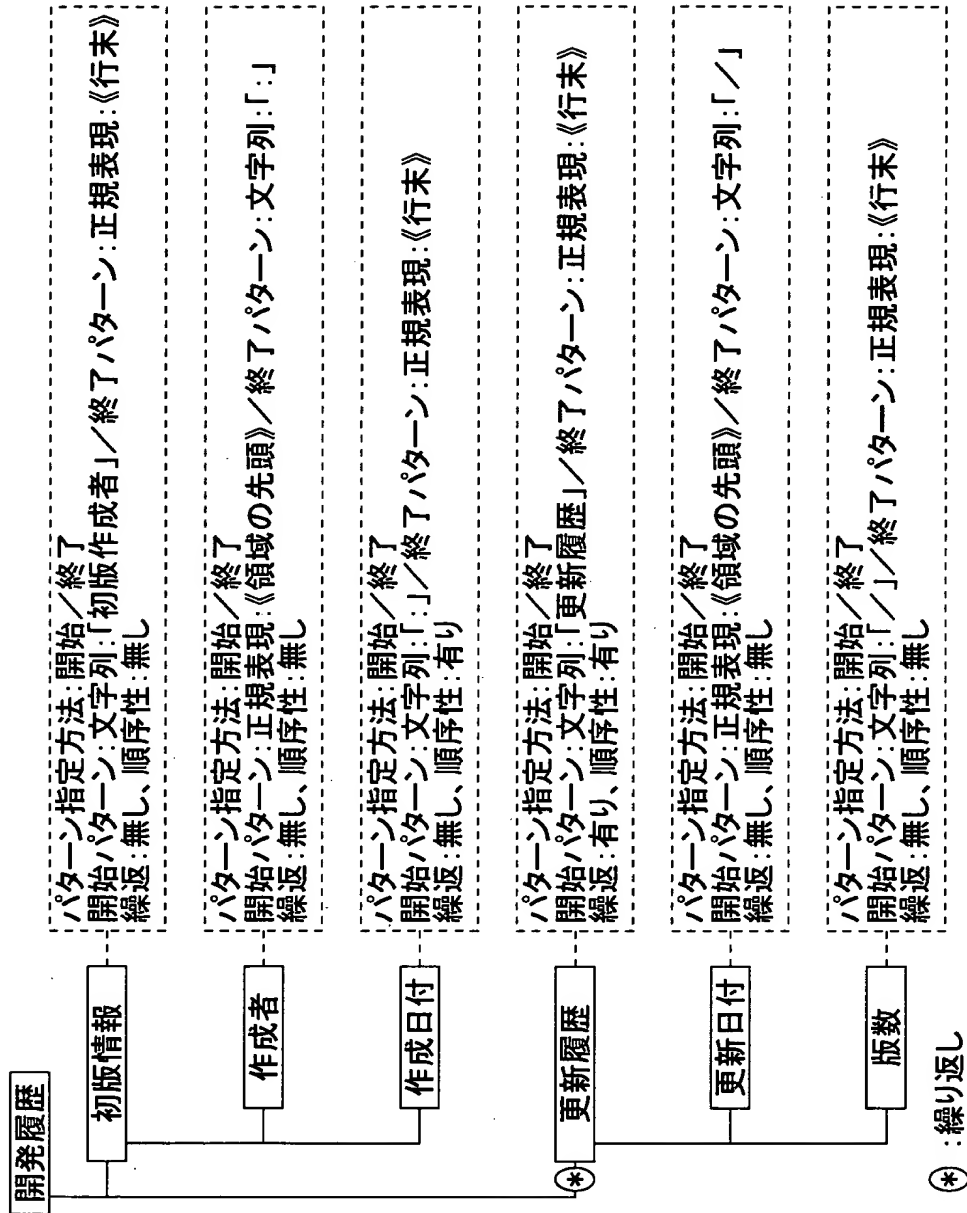
開始パターン: 文字列: 「/」

終了パターン: 正規表現: 《行末》

繰返: 無し

順序性: 無し

【図 20】



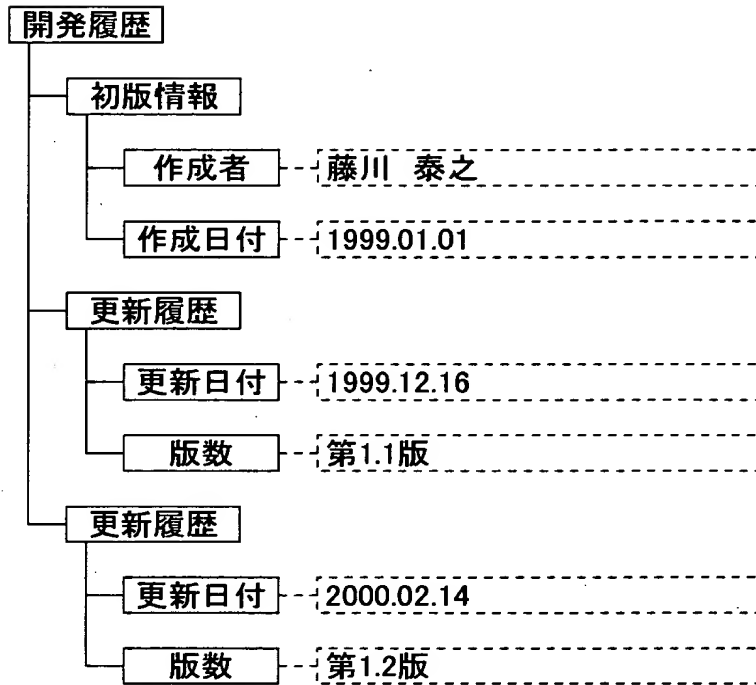
【図 21】

```

*****
初版作成者 藤川 泰之 : 1999.01.01
*****
更新履歴 1999.12.16 / 第1.1版
更新履歴 2000.02.14 / 第1.2版
*****

```

【図 2 2】

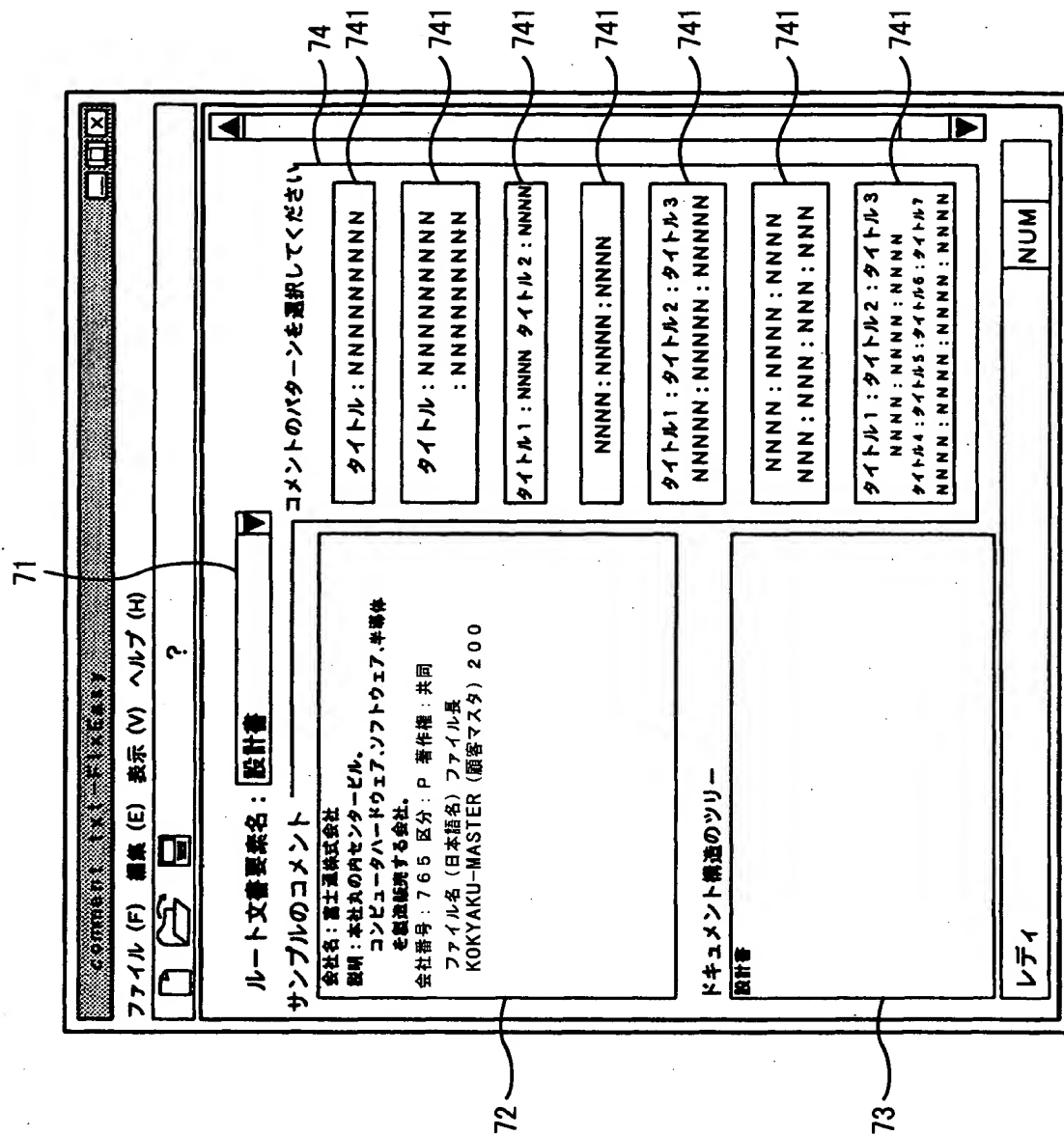


【図 2 3】

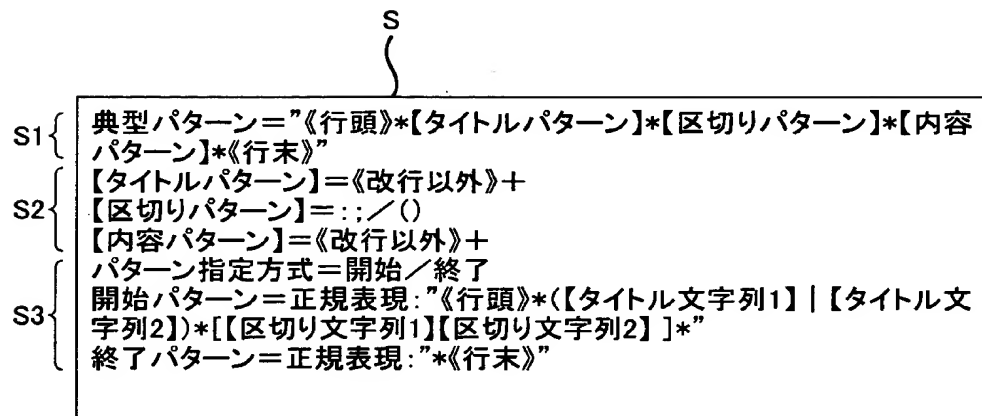
```

<開発履歴>
<初版情報>
<作成者>藤川 泰之</作成者>
<作成日付>1999.01.01</作成日付>
<更新履歴>
<更新日付>1999.12.16</更新日付>
<版数>第1.1版</版数>
</更新履歴>
<更新履歴>
<更新日付>2000.02.14</更新日付>
<版数>第1.2版</版数>
</更新履歴>
</開発履歴>
  
```

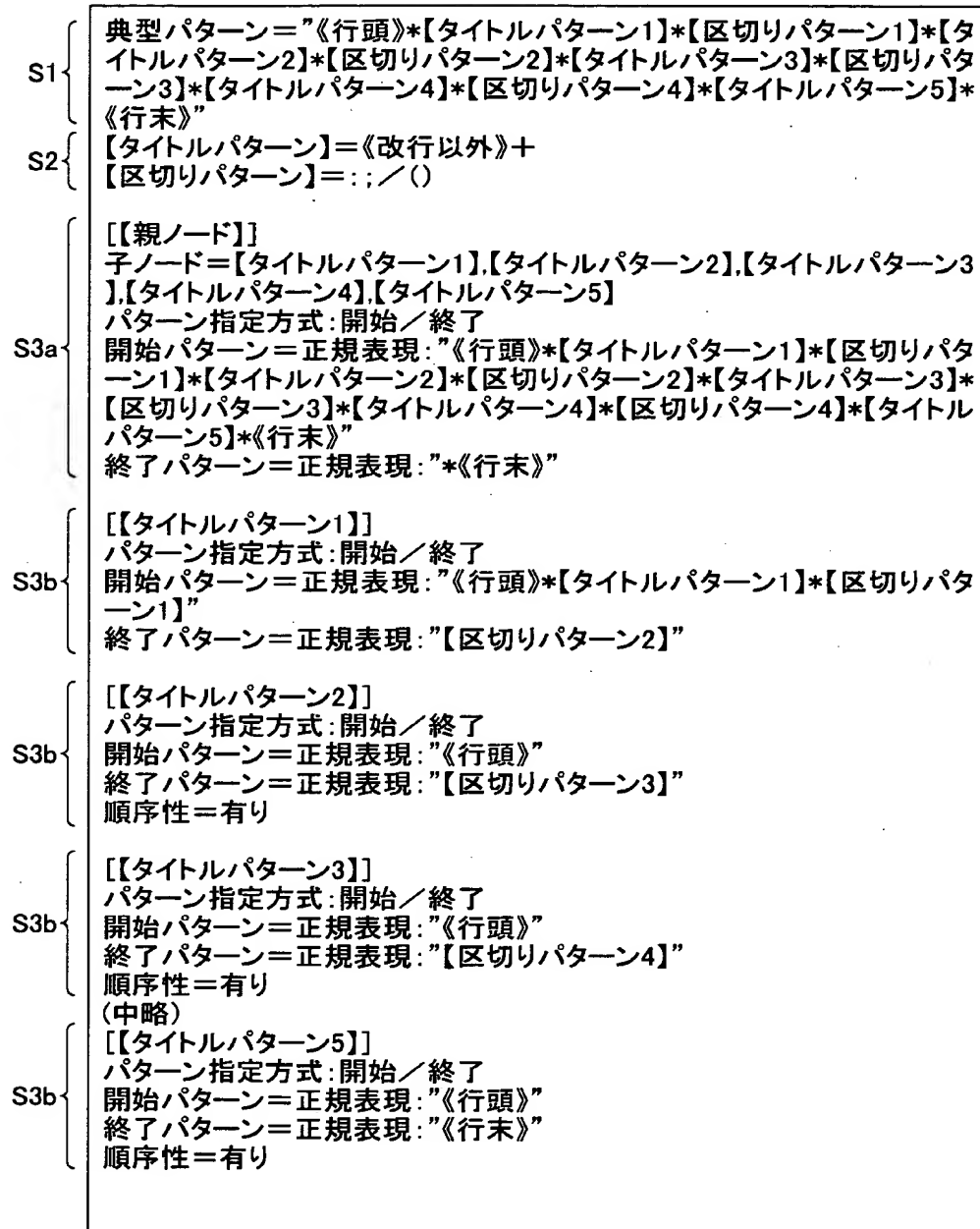
【図 24】



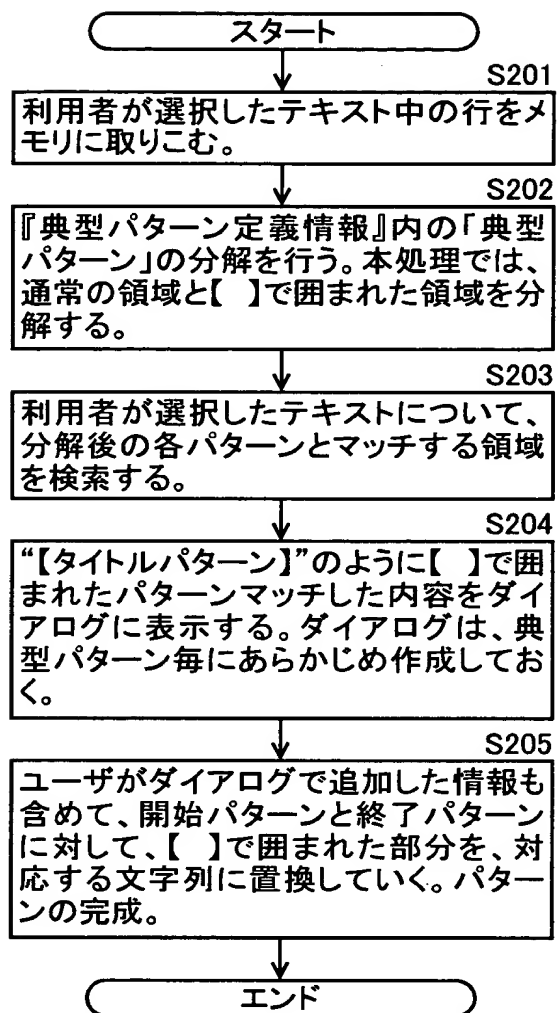
【図 2 5】



【図 2 6】



【図 2 7】



【図 28】

ファイル (F) 編集 (E) 表示 (V) ヘルプ (H)

?

ルート文書要素名: 設計書

サンプルのコメント

説明: 本社の内センタービル、
コンピュータハードウェア、ソフトウェア、半導体
を製造販売する会社。
会社番号: 765 区分: P 著作権: 共同
ファイル名 (日本語名) ファイル長
KOKYAKU-MASTER 履歴マスタ 200

コメントのパターンを選択してください

タイトル: NNNNNNNNNN

タイトル: NNNNNNNNNN
: NNNNNNNNNN

タイトル1: NNNN タイトル2: NNNN

タイトル: NNNNNNNNNN パターン

この部分の名前: 会社名

区切り文字列:

タイトル文字列:

会社名

区切り文字列:

☐ 繰り返し有り

追加

キャンセル

詳細設定

NNNN: NNNN: NNNN: NNNN

NUM

レディ

ドキュメント構造のツリー

設計書

701

704

705

702

703

700

706

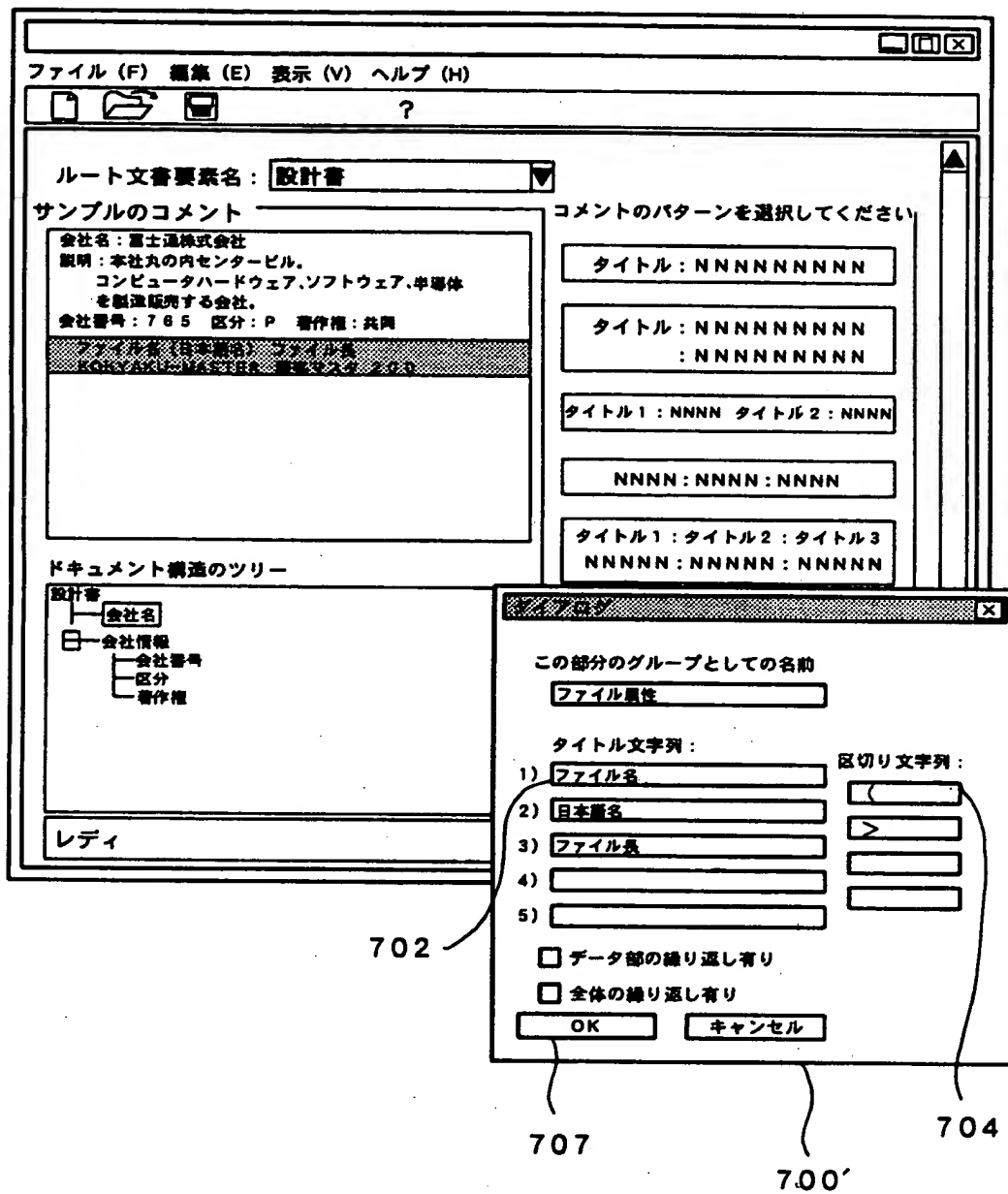
【图 29】

[会社名]
パターン指定方式:開始／終了
開始パターン＝正規表現:”《行頭》*(富士通)*[:]*”
終了パターン＝正規表現:”*《行末》”

【図 30】

COMMENT LIST - FIVE			
ファイル(F)	編集(E)	表示(V)	ヘルプ(H)
		?	
ルート文書要素名：設計書			
サンプルのコメント		コメントのパターンを選択してください	
会社名：富士通株式会社 説明：本社丸の内センタービル。 コンピュータハードウェア・ソフトウェア、半導体 を製造販売する会社。		タイトル：NNNNNNNNNN	
〒100-8555東京都千代田区外神田2-1-1 三井物産株式会社		タイトル：NNNNNNNNNN ：NNNNNNNNNN	
ファイル名（日本語名）ファイル長 KOKYAKU-MASTER（顧客マスタ）200		タイトル1：NNNN タイトル2：NNNN	
		NNNN：NNNN：NNNN	
ドキュメント構造のツリー		タイトル1：タイトル2：タイトル3 NNNN：NNNN：NNNN	
設計書 └─ 会社名		NNNN：NNNN：NNNN NNN：NNN：NNN：NNN	
		タイトル1：タイトル2：タイトル3 NNNN：NNNN：NNNN タイトル4：タイトル5：タイトル7 NNNN：NNNN：NNNN：NNNN	
レディ		NUM	

【図 3 1】



【図 3 2】

COMPANY TITLE FIXESYS ファイル (F) 編集 (E) 表示 (V) ヘルプ (H)		?	
ルート文書要案名: 設計書		コメントのボタンを選択してください	
サンプルのコメント		タイトル: NNNNNNNNNN タイトル: NNNNNNNNNN : NNNNNNNNNN タイトル1: NNNN タイトル2: NNNN NNNN: NNNN: NNNN タイトル1: タイトル2: タイトル3 NNNNN: NNNNN: NNNNN NNNN: NNNN: NNNN NNN: NNN: NNN タイトル1: タイトル2: タイトル3 NNNN: NNNN: NNNN タイトル4: タイトル5: タイトル6: タイトル7 NNNN: NNNN: NNNN: NNNN	
会社名: 富士通株式会社 住所: 本社丸の内センタービル コンピュータハードウェア・ソフトウェア、半導体 を製造販売する会社。 会社番号: 765 区分: P 著作権: 共同		ファイル名: 日本電気 ファイル名 KONANET-WASHER 番号: 265	
ドキュメント構造のツリー		設計書	
会社名 会社情報 会社番号 区分 著作権 ファイル属性 ファイル名 日本電気 ファイル名		会社名 会社情報 会社番号 区分 著作権 ファイル属性 ファイル名 日本電気 ファイル名	
レディ		NUM	

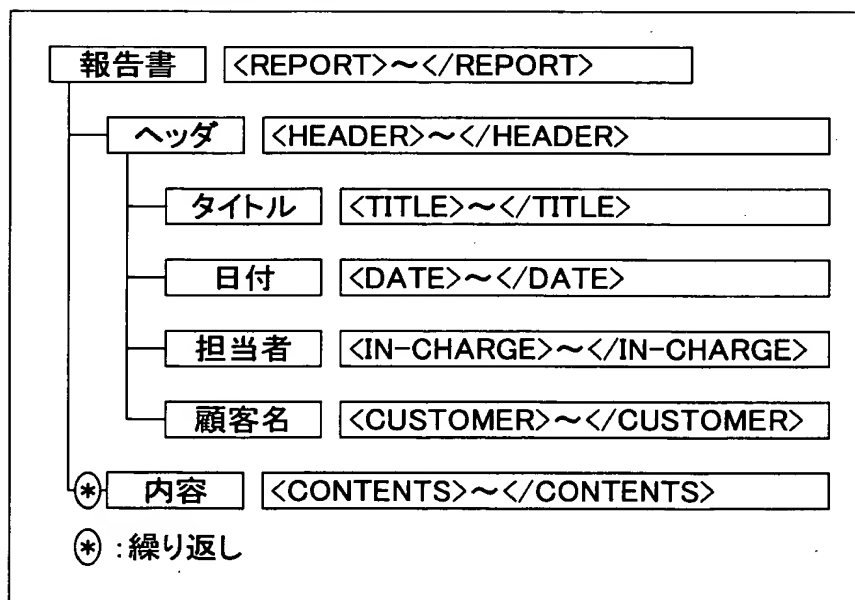
【図 3 3】

(a)
構造化テキスト(XML)の例

```

<REPORT>
<HEADER>
<TITLE>商談対応報告書</TITLE>
<DATE>1997.02.17</DATE>
<IN-CHARGE>藤川 泰之</IN-CHARGE>
<CUSTOMER>山田証券</CUSTOMER>
</HEADER>
<CONTENTS>
以下のSDASツールの打診あり。
・AA／BRMODELLER
・YPS
</CONTENTS>
<CONTENTS>
当社説明員によるデモを、来週中に
顧客先にて行う予定。同時に顧客の
QA票(開発元に送付済)の回答も行う。
</CONTENTS>
<REPORT>

```

(b)
文書構造定義(DTD)を図にしたもの

【書類名】 要約書

【要約】

【課題】 テキストから自動的に構造化文書を生成することができる構造化文書化システムを、提供する。

【解決手段】

D T D + パターンツリー作成部 5 1 は、D T D + パターン情報 R が定義する各要素の階層構造をツリーとして表すとともに、ツリーの各ノードに、該当する要素に関して指定された記述パターンを付加する。全体コントロール部 5 2 は、このツリーの各ノード毎に、指定された記述パターンに従った検索をパターン検索部 5 3 に依頼する。パターン検索部 5 3 は、指定された記述パターンに合致した領域を処理対象文書 T から抽出して、全体コントロール部 5 2 に回答する。全体コントロール部 5 2 は、各要素に対応するものとして回答されたテキストの領域の前後に、その要素に対応したタグを付加することにより、構造化文書 O を出力する。

【選択図】 図 3

認定・付加情報

特許出願の番号	特願 2000-027460
受付番号	50000125129
書類名	特許願
担当官	第七担当上席 0096
作成日	平成12年 2月 7日

<認定情報・付加情報>

【提出日】	平成12年 2月 4日
-------	-------------

出 願 人 履 歴 情 報

識別番号

[000005223]

1. 変更年月日 1996年 3月26日

[変更理由] 住所変更

住 所 神奈川県川崎市中原区上小田中4丁目1番1号
氏 名 富士通株式会社